User Manual for GUIDE ver. 41.1*

Wei-Yin Loh
Department of Statistics
University of Wisconsin–Madison
April 10, 2023

Contents

1 Warranty disclaimer 5

2 Introduction 6
  2.1 Installation 7
  2.2 \LaTeX 11

3 Program operation 12
  3.1 Required files 12
  3.2 Input file creation 16

4 Classification: RHC data 17
  4.1 Univariate splits 19
    4.1.1 Input file generation 19
    4.1.2 Contents of classin.txt 22
    4.1.3 Contents of classout.txt 23
    4.1.4 Contents of classfit.txt 31
    4.1.5 Contents of classpred.r 33
  4.2 Linear splits 35

*Based on work partially supported by grants from the U.S. Army Research Office, National Science Foundation, National Institutes of Health, Bureau of Labor Statistics, USDA Economic Research Service, and Eli Lilly & Co. Work on precursors to GUIDE additionally supported by IBM Research and Pfizer.
### Contents

1. **Input file generation**
   - 4.2.1 Input file generation .......................... 35
   - 4.2.2 Contents of `linearin.txt` ...................... 37
   - 4.2.3 Contents of `linearout.txt` .................... 38
   - 4.2.4 R code for plot ................................ 46

2. **Kernel discriminant models**
   - 4.3.1 Input file generation .......................... 47
   - 4.3.2 Contents of `ker2.out` ........................... 50

3. **Nearest-neighbor models**
   - 4.4.1 Input file generation .......................... 57
   - 4.4.2 Contents of `nn2.out` ............................ 59

4. **Missing-value flag variables: CE data**
   - 5.1 Classification tree .................................. 73
   - 5.1.1 Input file generation .......................... 73
   - 5.1.2 Contents of output file ......................... 75

5. **Least squares regression: CE data**
   - 6.1 Piecewise constant .................................. 84
   - 6.1.1 Input file creation .......................... 84
   - 6.1.2 Contents of `cons.out` ........................... 86
   - 6.2 Piecewise simple polynomial .......................... 91
   - 6.2.1 Input file creation .......................... 93
   - 6.2.2 Partial output ................................ 95
   - 6.2.3 Plots of data ................................ 99
   - 6.3 Stepwise linear .................................. 103
   - 6.3.1 Input file creation .......................... 103
   - 6.3.2 Results .................................. 105

6. **Quantile regression: CE data**
   - 7.1 Piecewise constant: one quantile .................. 114
   - 7.1.1 Input file creation .......................... 114
   - 7.2 Best simple linear .................................. 121
   - 7.2.1 Input file creation .......................... 121
   - 7.3 Two quantiles .................................. 130
   - 7.3.1 Input file creation .......................... 130
   - 7.3.2 Output file .................................. 133


Wei-Yin Loh 2 GUIDE manual
# Contents

13 Multiresponse: NMES data ........................................................................ 246
  13.1 Input file creation .............................................................................. 248
  13.2 Contents of `mult.out` ....................................................................... 250

14 Longitudinal response ........................................................................... 255
  14.1 Input file creation .............................................................................. 257
  14.2 Contents of `wage.out` ...................................................................... 259

15 Logistic regression .................................................................................. 265
  15.1 Piecewise constant ............................................................................ 266
    15.1.1 Input file creation ...................................................................... 266
    15.1.2 Contents of `logitc.out` .............................................................. 267
  15.2 Simple linear ...................................................................................... 271
    15.2.1 Input file creation ...................................................................... 271
    15.2.2 Contents of `logits.out` ............................................................... 273
  15.3 Multiple linear .................................................................................... 276
    15.3.1 Input file creation ...................................................................... 276
    15.3.2 Results ....................................................................................... 278

16 Importance scoring .................................................................................. 285
  16.1 Classification: RHC data ................................................................. 285
    16.1.1 Input file creation ...................................................................... 285
    16.1.2 Contents of `imp.out` ................................................................. 286
  16.2 Censored response with R variable .................................................... 293
    16.2.1 Input file creation ...................................................................... 293
    16.2.2 Partial contents of `imp_surv.out` ............................................. 295

17 Causal inference ...................................................................................... 296
  17.1 Input file creation .............................................................................. 298
  17.2 Contents of `prop30.out` ................................................................. 299

18 Differential item functioning .................................................................. 307

19 Bootstrap confidence intervals ................................................................ 312

20 Tree ensembles ....................................................................................... 315
  20.1 GUIDE forest: CE data ................................................................. 318
    20.1.1 Input file creation ...................................................................... 318
    20.1.2 Contents of `gf.out` ................................................................. 320
Redistribution and use in binary forms, with or without modification, are permitted provided that the following condition is met:

Redistributions in binary form must reproduce the above copyright notice, this condition and the following disclaimer in the documentation and/or other materials provided with the distribution.

THIS SOFTWARE IS PROVIDED BY WEI-YIN LOH “AS IS” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL WEI-YIN LOH BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

The views and conclusions contained in the software and documentation are those of the author and should not be interpreted as representing official policies, either expressed or implied, of the University of Wisconsin.
2 Introduction

GUIDE stands for Generalized, Unbiased, Interaction Detection and Estimation. It is an algorithm for construction of classification and regression trees and forests. It is a descendent of the FACT (Loh and Vanichsetakul, 1988), SUPPORT (Chaudhuri et al., 1994, 1995), QUEST (Loh and Shih, 1997), CRUISE (Kim and Loh, 2001, 2003), and LOTUS (Chan and Loh, 2004; Loh, 2006a) algorithms. GUIDE is the only classification and regression tree algorithm with all these features:

1. Unbiased variable selection with and without missing data.
2. Unbiased importance scoring and thresholding of predictor variables.
3. Automatic handling of missing values without requiring prior imputation.
4. Provision for multiple missing-value codes and missing-value flag variables.
5. Optional automatic creation of missing-value indicator variables for regression.
6. Periodic or cyclic variables, such as angular direction, hour of day, day of week, month of year, and seasons.
7. Subgroup identification for differential treatment effects.
8. Linear splits and kernel and nearest-neighbor node models for classification trees.
9. Weighted least squares, least median of squares, logistic, quantile, Poisson, and relative risk (proportional hazards) regression models.
10. Univariate, multivariate, censored, and longitudinal response variables.
11. Pairwise interaction detection at each node.
12. Categorical variables for splitting only, fitting only (via 0-1 dummy variables), or both in regression tree models.
13. Tree ensembles (bagging and forests).

Tables 1 and 2 compare the features of GUIDE with QUEST, CRUISE, C4.5 (Quinlan, 1993), CTREE (Hothorn et al., 2006), MOB (Hothorn and Zeileis, 2015), RPART (Therneau et al., 2017) \(^1\), and M5' (Quinlan, 1992; Witten and Frank, 2000).

\(^1\)RPART is an implementation of CART (Breiman et al., 1984) in R. CART is a registered trademark of California Statistical Software, Inc.

2.1 Installation


Linux: There are two 64-bit executables to choose from: Intel or gfortran. Both versions are compiled in Ubuntu 20.0. Unzip the file with “gunzip guide.gz” and, if necessary, make it executable by typing “chmod a+x guide” in a Terminal window. To execute, type “./guide”.

macOS: There are five versions to choose from. The NAG versions do not require additional software to be installed; the gfortran versions require Xcode as explained below. Download the desired guide.gz file and double-click it to gunzip. The make it executable by typing the command “chmod a+x guide” in a Terminal application in the folder where the file is located. If this still does not allow you to run the app, carry out these steps:

1. In the Finder on your Mac, locate the file guide.
2. Control-click the guide icon, then choose Open from the shortcut menu.
3. Click Open.

Now you can start the program by typing “./guide” in the Terminal window where the file guide resides.

NAG Fortran for Apple M1. This is the only native version for M1 Macs. It is compiled with macOS Monterey 12.4.

NAG Fortran for Intel. This is version is for Intel Macs, but it works for M1 Macs too. It is compiled with macOS Catalina 10.15.7 but also works on Big Sur.

https://github.com/fxcoudert/gfortran-for-macOS/releases
Table 1: Comparison of GUIDE, QUEST, CRUISE, CART, C4.5, and CTREE classification tree algorithms. Node models: S = simple, K = kernel, L = linear discriminant, N = nearest-neighbor.

<table>
<thead>
<tr>
<th></th>
<th>GUIDE</th>
<th>QUEST</th>
<th>CRUISE</th>
<th>RPART</th>
<th>C4.5</th>
<th>CTREE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unbiased splits</td>
<td>Yes</td>
<td>Yes if no missing values</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes if no missing values</td>
</tr>
<tr>
<td>Splits per node</td>
<td>2</td>
<td>2</td>
<td>≥ 2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Linear splits</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Categorical variable splits</td>
<td>Subsets</td>
<td>Subsets</td>
<td>Subsets</td>
<td>Subsets</td>
<td>Atoms</td>
<td>Subsets</td>
</tr>
<tr>
<td>Periodic variable splits</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Interaction tests</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Class priors</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Misclassification costs</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No^a</td>
</tr>
<tr>
<td>Case weights</td>
<td>No^b</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes^c</td>
</tr>
<tr>
<td>Node models</td>
<td>S, K, N</td>
<td>S</td>
<td>S, L</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>Missing values in splits</td>
<td>Missing as observed</td>
<td>Node mean or mode imputation</td>
<td>Surrogate splits</td>
<td>Surrogate splits</td>
<td>Weights</td>
<td>Random splits^d</td>
</tr>
<tr>
<td>Missing-value flag variables</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Pruning</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Tree diagrams</td>
<td>Text and \LaTeX</td>
<td>R</td>
<td>Text</td>
<td>R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bagging</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Forests</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>cforest</td>
</tr>
<tr>
<td>Importance scores</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

^auser defined
^bpositive weights treated as 1
^cnon-negative integer counts
^dsurrogate splits is a non-default option
Table 2: Comparison of GUIDE, RPART, M5’, and MOB regression tree algorithms

<table>
<thead>
<tr>
<th>Feature</th>
<th>GUIDE</th>
<th>RPART</th>
<th>M5’</th>
<th>MOB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unbiased splits</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Interaction tests</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Loss functions</td>
<td>Weighted least squares, least median of squares, logistic, quantile, Poisson, least absolute deviations</td>
<td>Least squares, least squares</td>
<td>Generalized linear models</td>
<td></td>
</tr>
<tr>
<td>Censored response</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Longitudinal and multi-response</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Node models</td>
<td>Constant, multiple, step-wise linear, polynomial, ANCOVA</td>
<td>Constant, stepwise</td>
<td>Constant, multiple linear</td>
<td></td>
</tr>
<tr>
<td>Variable roles</td>
<td>Split only, fit only, both, neither, weight, offset</td>
<td>Split only</td>
<td>Split and fit</td>
<td>Similar to GUIDE</td>
</tr>
<tr>
<td>Categorical variable splits</td>
<td>Subsets</td>
<td>Subsets</td>
<td>Atomic</td>
<td>Subsets</td>
</tr>
<tr>
<td>Periodic variables</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Tree diagrams</td>
<td>Text and \LaTeX</td>
<td>R</td>
<td>PostScript</td>
<td>R</td>
</tr>
<tr>
<td>Sampling weights</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No\textsuperscript{a}</td>
</tr>
<tr>
<td>Transformations</td>
<td>Powers and products</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Missing values in splits</td>
<td>Missing as observed</td>
<td>Surrogate splits</td>
<td>Mean/mode imputation</td>
<td>Random splits</td>
</tr>
<tr>
<td>Missing values in linear predictors</td>
<td>Node mean imputation &amp; missing-value indicators</td>
<td>N/A</td>
<td>Global imputation</td>
<td>Omitted</td>
</tr>
<tr>
<td>Missing-value flag variables</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Bagging &amp; forests</td>
<td>Yes &amp; yes</td>
<td>No &amp; no</td>
<td>No &amp; no</td>
<td>cforest</td>
</tr>
<tr>
<td>Importance scores</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes\textsuperscript{b}</td>
</tr>
</tbody>
</table>

\textsuperscript{a}replicate weights only
\textsuperscript{b}from cforest or ctree
Monterey 12.6.2 with gfortran 12.1. This version is compiled with Xcode 14.2.
3. Double-click the disk image to install gfortran.

Big Sur 11.7.1 with gfortran 11.2. This version is compiled with Xcode 13.2.1.
3. Double-click the disk image to install gfortran.

gfortran for High Sierra. This version is for Mac that cannot be upgrade above macOS High Sierra. It is compiled with Xcode 10.1 and gfortran 5.1. Follow these steps to ensure that the gfortran libraries are placed in the right place:
3. Open a Terminal window and type (or copy and paste):
   (a) cd ~/Downloads
   (b) gunzip gcc-5.1-bin.tar.gz
   (c) sudo tar -xvf gcc-5.1-bin.tar -C /

Windows: There are two 64-bit executables to choose from: gfortran or Intel. Both versions are compiled under Windows 10. The gfortran version is preferable because the Intel version sometimes crashes due to a compiler bug. Download the file guide.zip and unzip it (right-click on file icon and select “Extract all”). The resulting file guide.exe may be placed in one of three places:
1. Top level of your C drive. Type “C:\guide” in a Command Prompt window to execute—see Section 3.1.
2. A folder that contains your data files. Type “guide” in that folder to execute.
3. A folder on your search path. Type “guide” anywhere to execute.
2.2 \LaTeX

GUIDE uses the public-domain software \LaTeX \r{ } (\url{http://www.ctan.org}) to produce tree diagrams. The \LaTeX \r{ } software may be obtained from:

**Linux:** TeX Live \url{http://www.tug.org/texlive/}

**Mac:** MacTeX \url{http://tug.org/mactex/} or MikTeX \url{https://miktex.org/howto/install-miktex-mac}. Both include the TeXShop GUI frontend.

**Windows:** MikTeX \url{https://miktex.org/howto/install-miktex} or proTeXt \url{http://www.tug.org/protext/}. The former includes the TeXShop GUI frontend and latter includes TeXStudio.

The \LaTeX \r{ } files produced by GUIDE can be edited to change colors, node sizes, etc., in the trees; see \textit{pstricks manual} \r{ } (\url{http://tug.org/PSTricks/main.cgi/}). There are two ways to generate pdf figures of the tree diagrams. In the following, assume that the \LaTeX \r{ } file is named \texttt{diagram.tex}.  

1. **Terminal window (simplest).** Type these three commands in the Terminal (Linux or Mac) or Command Prompt (Win) window where the \LaTeX \r{ } file (say, \texttt{diagram.tex}) was produced.

   (a) \texttt{latex diagram}

   (b) \texttt{dvips diagram}

   (c) \texttt{ps2pdf diagram.ps}

   The first command produces a file called \texttt{diagram.dvi}. The second command converts the latter to postscript file called \texttt{diagram.ps} (which can be edited with any postscript app). The third command turns it into a pdf file with name \texttt{diagram.pdf}.

2. **TeXShop, TeXworks, or TeXStudio.** Double-click \texttt{diagram.tex} to load it into one of these apps. Select XeLaTeX to typeset it to pdf.

   In Mac OSX, the Preview app can open postscript and pdf files for conversion to jpg, png, and other formats. In Windows, the same can be done with ImageMagick \r{ } (\url{https://www.imagemagick.org/}). For inclusion of the pdf figures in MS PowerPoint or Word documents, convert them to jpg for Mac OSX and png for Windows.

Wei-Yin Loh 11 GUIDE manual
3 Program operation

GUIDE runs within a terminal window of the computer operating system.

**Linux.** Any terminal program will do.

**Mac OSX.** The program is called **Terminal**; it is in the **Applications Folder**.

**Windows.** The terminal program is started from the **Start button** by choosing **All Programs → Accessories → Command Prompt**

Do not double-click the GUIDE icon on the desktop!

After the terminal window is opened, change to the folder where the data and program files are stored. Mac and Windows users are unfamiliar with terminal commands may consult https://wiredpen.com/resources/basic-unix-commands-for-osx/ and https://cmdref.net/os/windows/command/index.html, respectively.

3.1 Required files

GUIDE requires two text files to begin.

**Data file:** This file contains the data from the training sample. Each data record consists of observations on the dependent variable, the predictor (i.e., $X$ or independent) variables, and optional weight, missing value flag, time, offset, periodic, and event indicator (for censored responses) variables. Entries in each record are comma, space, or tab delimited (multiple spaces are treated as one space, but not for commas). A record can occupy more than one line in the file, but each record must begin on a new line.

Values of categorical variables can contain any ascii character except single and double quotation marks, which are used to enclose values that contain spaces and commas. Values can be up to 60 characters long. Class labels are truncated to 10 characters in tabular output.

A common problem among first-time users is getting the data file in proper shape. If the data are in a spreadsheet and there are no empty cells, export them to a **MS-DOS Comma Separated** (csv) file (the MS-DOS CSV format takes care of carriage return and line feed characters properly). If there are empty cells, a good solution is to read the spreadsheet into R (using `read.csv`
with proper specification of the `na.strings` argument), verify that the data are correctly read, and then export them to a text file using either `write.table` or `write.csv`.

**Note to R users**: GUIDE can optionally generate R code for the tree model and its prediction function. Because GUIDE treats "NA" (with quotes) the same as `NA` (without quotes), the two are treated as missing values in the R function.

**Description file**: This provides information about the name and location of the data file, column locations and names of the variables, and their roles in the analysis. Different models may be fitted by changing the roles of the variables. An example description file is `rhcdsc1.txt` whose contents follow.

```
rhcdscl.txt
NA
2
1 X x
2 cat1 c
3 cat2 c
4 ca c
5 sadmdte x
6 dschdte x
7 dthdte x
8 lstctdte x
9 death x
10 cardiohx c
11 chfhx c
12 dementhx c
13 psychhx c
14 chrpulhx c
15 renalhx c
16 liverhx c
17 gibledhx c
18 malighx c
19 immunhx c
20 transhx c
21 amihx c
22 age n
```
3.1 Required files

23 sex c
24 edu n
25 surv2md1 n
26 das2d3pc n
27 t3d30 x
28 dth30 x
29 aps1 n
30 scoma1 n
31 meanbp1 n
32 wblc1 n
33 hrt1 n
34 resp1 n
35 temp1 n
36 pafii1 n
37 alb1 n
38 hema1 n
39 bili1 n
40 crea1 n
41 sod1 n
42 pot1 n
43 paco21 n
44 ph1 n
45 swang1 d
46 wtkilo1 n
47 dnr1 c
48 mnsclas c
49 resp c
50 card c
51 neuro c
52 gastr c
53 renal c
54 meta c
55 hema c
56 seps c
57 trauma c
58 ortho c
59 adld3p n
60 urini1 n
The 1st line gives the name of the data file. If the file is not in the current folder, its full path must be given (e.g., "c:\data\rhndata.txt" for Windows users or "~/Data/rhndata.txt" for Mac users) surrounded by matching quotes (because it contains non-alphanumeric characters). The 2nd line gives the missing value code, which can be up to 80 characters long. If it contains non-alphanumeric characters, it too must be surrounded by matching quotation marks. A missing value code must appear in the second line of the file even if there are no missing values in the data (in which case any character string not present among the data values can be used). The 3rd line gives the line number of the first data record in the data file. A “2” is shown here because the variable names appear in the first line of rhndata.txt. If the 1st line of the data file contains the 1st record, this entry would be “1”. Blank lines in the data and description files are ignored. The column location, name and role of each variable comes next (in that order), with one line for each variable.

Variable names must begin with an alphabet and be not more than 60 characters long. If a name contains non-alphanumeric characters, it must be enclosed in matching single or double quotes. Spaces and the four special characters, #, %, {, and }, in a variable name are replaced by dots (periods) in the outputs. Variable names are truncated to 10 characters in tabular text output (but not in R output). Leading and trailing spaces in variable names are dropped.

The letters (lower or upper case) below are the permissible roles:

b Categorical variable used both for splitting and for node modeling in regression. Such variables are converted to 0-1 dummy variables when fitting models within nodes for regression. They are converted to c type for classification.

c Categorical variable used for splitting only.

d Dependent variable or death indicator variable. Except for longitudinal and multiple response data (Sec. 13), there can only be one d variable. For censored responses in proportional hazards models, it is the 0-1 event (death) indicator. For all other models, it is the response variable. It can take character string values for classification.
3.2 Input file creation

GUIDE is started by typing its (lowercase) name in a terminal and then typing “1” to answer some questions and save the answers into a file. In the following, the sign (>) is the computer prompt (not to be typed!).

Wei-Yin Loh 16 GUIDE manual
Table 3: Predictor variable role descriptors

<table>
<thead>
<tr>
<th>Type of variable</th>
<th>Role of variable</th>
<th>Split nodes</th>
<th>Fit node models</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical</td>
<td></td>
<td>c</td>
<td>i</td>
<td>b</td>
</tr>
<tr>
<td>Numerical</td>
<td></td>
<td>s</td>
<td>f</td>
<td>n</td>
</tr>
</tbody>
</table>

Doctors believe that direct measurement of cardiac function by right heart catheterization (RHC) is beneficial for some critically ill patients. The file `rhcdata.txt` contains observations on more than 60 variables for 5735 patients from 5 medical centers over 5 years (Connors et al., 1996). The variable `swang1` takes values “RHC” and “NoRHC”, indicating whether or not a patient received RHC. Variable `dth30` is 1 if death occurs within 30 days of hospital admission and 0 otherwise; `death` is 1 if the subject eventually dies and 0 if death is unknown. Other variables are given in Tables 4–7.

To construct a classification tree for predicting `swang1`, we need to generate an input file from the description file `rhdsc1.txt`, which specifies `swang1` as a `d` variable and `dth30` and `death` both as `x`. When GUIDE prompts for a selection, there is usually range of permissible values given within square brackets and a default choice (indicated by the symbol `<cr>`=). The default may be selected by pressing the ENTER or RETURN key.
### Table 4: RHC demographic & outcome variables [missing values in brackets]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>swang1</td>
<td>Right heart catheterization (RHC)</td>
<td>0</td>
</tr>
<tr>
<td>age</td>
<td>Age in years</td>
<td>0</td>
</tr>
<tr>
<td>sex</td>
<td>Sex (female/male)</td>
<td>0</td>
</tr>
<tr>
<td>wtkilo1</td>
<td>Weight in kilograms</td>
<td>515</td>
</tr>
<tr>
<td>edu</td>
<td>Years of Education</td>
<td>0</td>
</tr>
<tr>
<td>race</td>
<td>Race</td>
<td>0</td>
</tr>
<tr>
<td>income</td>
<td>Income bracket (&lt;11k, 11–25k, 25–50k, &gt;50k)</td>
<td>0</td>
</tr>
<tr>
<td>ninsclas</td>
<td>Medical insurance (Medicaid, Medicare, Medicare &amp; Medicaid, no insurance, private, private &amp; Medicare)</td>
<td>0</td>
</tr>
<tr>
<td>t3d30</td>
<td>Days from admission to death within 30 days</td>
<td>0</td>
</tr>
<tr>
<td>dth30</td>
<td>Death indicator for t3d30 (0=no, 1=yes)</td>
<td>0</td>
</tr>
<tr>
<td>survtime</td>
<td>Days from admission to death or last contact day</td>
<td>0</td>
</tr>
<tr>
<td>death</td>
<td>Death indicator for survtime (0=no, 1=yes)</td>
<td>0</td>
</tr>
<tr>
<td>transhx</td>
<td>Transfer (&gt; 24 hours) from another hospital (no/yes)</td>
<td>0</td>
</tr>
</tbody>
</table>

### Table 5: RHC disease variables [missing values in brackets]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat1</td>
<td>Primary disease category (9 levels)</td>
<td>0</td>
</tr>
<tr>
<td>cat2</td>
<td>Secondary disease category (6 levels)</td>
<td>2798</td>
</tr>
<tr>
<td>ca</td>
<td>Cancer (3 levels)</td>
<td>0</td>
</tr>
<tr>
<td>card</td>
<td>Cardiovascular diagnosis</td>
<td>0</td>
</tr>
<tr>
<td>gastr</td>
<td>Gastrointestinal diagnosis</td>
<td>0</td>
</tr>
<tr>
<td>hema</td>
<td>Hematologic diagnosis</td>
<td>0</td>
</tr>
<tr>
<td>meta</td>
<td>Metabolic diagnosis</td>
<td>0</td>
</tr>
<tr>
<td>neuro</td>
<td>Neurological diagnosis</td>
<td>0</td>
</tr>
<tr>
<td>ortho</td>
<td>Orthopedic diagnosis</td>
<td>0</td>
</tr>
<tr>
<td>renal</td>
<td>Renal diagnosis</td>
<td>0</td>
</tr>
<tr>
<td>resp</td>
<td>Respiratory diagnosis</td>
<td>0</td>
</tr>
<tr>
<td>seps</td>
<td>Sepsis diagnosis</td>
<td>0</td>
</tr>
<tr>
<td>trauma</td>
<td>Trauma diagnosis</td>
<td>0</td>
</tr>
</tbody>
</table>
4.1 Univariate splits

The default classification tree employs only one variable to split each node. We demonstrate this first.

4.1.1 Input file generation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: classin.txt
   Input 1 for model fitting, 2 for importance or DIF scoring,
   3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: classout.txt
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1):
   Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
   Input name of data description file (max 100 characters);
   enclose with matching quotes if it has spaces: rhcdsc1.txt
   Reading data description file ...
   Training sample file: rhcdata.txt
   Missing value code: NA
   Records in data file start on line 2

Table 6: RHC medical history variables [#missing values in brackets]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>amihx</td>
<td>Definite myocardial infarction (no/yes) [0]</td>
<td></td>
</tr>
<tr>
<td>cardiohx</td>
<td>Acute MI, peripheral vascular disease, severe cardiovascular symptoms [0]</td>
<td></td>
</tr>
<tr>
<td>chfhx</td>
<td>Congestive heart failure (no/yes) [0]</td>
<td></td>
</tr>
<tr>
<td>chrpulhx</td>
<td>Chronic or severe pulmonary disease (no/yes) [0]</td>
<td></td>
</tr>
<tr>
<td>dementhx</td>
<td>Dementia, stroke or cerebral infarction, Parkinson’s disease (no/yes) [0]</td>
<td></td>
</tr>
<tr>
<td>gibledhx</td>
<td>Upper GI bleeding (no/yes) [0]</td>
<td></td>
</tr>
<tr>
<td>liverhx</td>
<td>Cirrhosis, hepatic failure (no/yes) [0]</td>
<td></td>
</tr>
<tr>
<td>malighx</td>
<td>Solid tumor, metastatic disease, chronic leukemia/myeloma, acute leukemia, lymphoma (no/yes) [0]</td>
<td></td>
</tr>
<tr>
<td>immunhx</td>
<td>Immunosuppression, organ transplant, HIV positivity, diabetes mellitus, connective tissue disease (no/yes) [0]</td>
<td></td>
</tr>
<tr>
<td>psychhx</td>
<td>Psychiatric history, active psychosis or severe depression (no/yes) [0]</td>
<td></td>
</tr>
<tr>
<td>renalhx</td>
<td>Chronic renal disease, chronic hemodialysis or peritoneal dialysis (no/yes) [0]</td>
<td></td>
</tr>
</tbody>
</table>
Table 7: RHC admission variables [#missing values in brackets]; PaO2 is partial pressure of arterial oxygen, FiO2 is fraction of inspired oxygen

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>alb1</td>
<td>Albumin [0]</td>
<td></td>
</tr>
<tr>
<td>bili1</td>
<td>Bilirubin [0]</td>
<td></td>
</tr>
<tr>
<td>crea1</td>
<td>Serum creatinine [0]</td>
<td></td>
</tr>
<tr>
<td>hema1</td>
<td>Hematocrit [0]</td>
<td></td>
</tr>
<tr>
<td>hrt1</td>
<td>Heart rate [159]</td>
<td></td>
</tr>
<tr>
<td>meanbp1</td>
<td>Mean blood pressure [80]</td>
<td></td>
</tr>
<tr>
<td>pot1</td>
<td>Serum potassium [0]</td>
<td></td>
</tr>
<tr>
<td>pafi1</td>
<td>PaO2/(0.01*FiO2) [0]</td>
<td></td>
</tr>
<tr>
<td>paco21</td>
<td>Partial pressure of arterial carbon dioxide [0]</td>
<td></td>
</tr>
<tr>
<td>ph1</td>
<td>Serum ph [0]</td>
<td></td>
</tr>
<tr>
<td>resp1</td>
<td>Respiration rate [136]</td>
<td></td>
</tr>
<tr>
<td>scoma1</td>
<td>Glasgow coma score [0]</td>
<td></td>
</tr>
<tr>
<td>sod1</td>
<td>Serum sodium [0]</td>
<td></td>
</tr>
<tr>
<td>temp1</td>
<td>Temperature (Celsius) [0]</td>
<td></td>
</tr>
<tr>
<td>urin1</td>
<td>Urine output [3028]</td>
<td></td>
</tr>
<tr>
<td>wblc1</td>
<td>White blood cell count [0]</td>
<td></td>
</tr>
<tr>
<td>aps1</td>
<td>APACHE III score ignoring coma [0]</td>
<td></td>
</tr>
<tr>
<td>addld3p</td>
<td>Katz Activities of Daily Living Scale [3016]</td>
<td></td>
</tr>
<tr>
<td>das2d3pc</td>
<td>DASI (Duke Activity Status Index) [0]</td>
<td></td>
</tr>
<tr>
<td>dnr1</td>
<td>DNR (do-not-resuscitate) status [0]</td>
<td></td>
</tr>
<tr>
<td>surv2md1</td>
<td>Estimated probability of 2-month survival [0]</td>
<td></td>
</tr>
</tbody>
</table>
4.1 Univariate splits

23 N variables changed to S
D variable is swang1
Reading data file ...
Number of records in data file: 5735
Length of longest entry in data file: 19
Checking for missing values ...
Finished checking
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Recoding D values to integers
Finished recoding
Number of classes: 2
Assigning integer codes to values of 30 categorical variables
Finished assigning codes to 10 categorical variables
Finished assigning codes to 20 categorical variables
Finished assigning codes to 30 categorical variables
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...
Finished processing 5000 of 5735 observations
Data checks complete
Creating missing value indicators ...
Rereading data ...
Class #Cases Proportion
NoRHC 3551 0.61918047
RHC 2184 0.38081953
Total #cases w/ #missing
#cases miss. D ord. vals #X-var #N-var #F-var #S-var
5735 0 5157 10 0 0 23
#P-var #M-var #B-var #C-var #I-var
0 0 0 30 0
Number of cases used for training: 5735
Number of split variables: 53
Number of cases excluded due to 0 W or missing D variable: 0
Finished reading data file
Choose 1 for estimated priors, 2 for equal priors, 3 to input priors from a file
Input 1, 2, or 3 ([1:3], <cr>=1):

Choose 1 for unit misclassification costs, 2 to input costs from a file
Input 1 or 2 ([1:2], <cr>=1):
Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):
Input file name to store LaTeX code (use .tex as suffix): class.tex
You can store the variables and/or values used to split and fit in a file
Choose 1 to skip this step, 2 to store split and fit variables, 3 to store split variables and their values
4.1 Univariate splits

Input your choice ([1:3], <cr>=1):
Input 2 to save fitted values and node IDs, 1 otherwise ([1:2], <cr>=2):
Input name of file to store node ID and fitted value of each case: classfit.txt
Input 2 to write R function for predicting new cases, 1 otherwise ([1:2], <cr>=2):
Input file name: classpred.r
Input rank of top variable to split root node ([1:53], <cr>=1):

Input file is created!
Run GUIDE with the command: guide < classin.txt

4.1.2 Contents of classin.txt

The resulting input file is given below. Each line contains a value followed by all the permissible values in parentheses. GUIDE reads only the first value in each row.

GUIDE (do not edit this file unless you know what you are doing)
41.0 (version of GUIDE that generated this file)
1 (1=model fitting, 2=importance or DIF scoring, 3=data conversion)
"classout.txt" (name of output file)
1 (1=one tree, 2=ensemble)
1 (1=classification, 2=regression, 3=propensity score grouping)
1 (1=simple model, 2=nearest-neighbor, 3=kernel)
1 (0=linear 1st, 1=univariate 1st, 2=skip linear, 3=skip linear and interaction)
1 (0=tree with fixed no. of nodes, 1=prune by CV, 2=by test sample, 3=no pruning)
"rhcdscl.txt" (name of data description file)
10 (number of cross-validations)
0.250 (SE number for pruning)
1 (1=estimated priors, 2=equal priors, 3=other priors)
1 (1=unit misclassification costs, 2=other)
2 (1=split point from quantiles, 2=use exhaustive search)
1 (1=default max. number of split levels, 2=specify no. in next line)
1 (1=default min. node size, 2=specify min. value in next line)
2 (0=no LaTeX code, 1=tree without node numbers, 2=tree with node numbers)
"class.tex" (latex file name)
1 (1=color terminal nodes, 2=no colors)
2 (0=#errors, 1=sample sizes, 2=sample proportions, 3=posterior probs, 4=nothing)
1 (1=no storage, 2=store fit and split variables, 3=store split variables and values)
2 (1=do not save fitted values and node IDs, 2=save in a file)
"classfit.txt" (file name for fitted values and node IDs)
2 (1=do not write R function, 2=write R function)
"classpred.r" (R code file)
1 (rank of top variable to split root node)
4.1 Univariate splits

4.1.3 Contents of classout.txt

The classification tree model is obtained by executing the command “guide < classin.txt” in the terminal window. The output file classout.txt, with annotations in blue, follow.

Classification tree
Pruning by cross-validation
Data description file: rhcdsc1.txt name of description file
Training sample file: rhcdata.txt name of data file
Missing value code: NA
Records in data file start on line 2
23 N variables changed to S
D variable is swang1
Number of records in data file: 5735
Length of longest entry in data file: 19
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Number of classes: 2
Training sample class proportions of D variable swang1:

<table>
<thead>
<tr>
<th>Class</th>
<th>#Cases</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoRHC</td>
<td>3551</td>
<td>0.61918047</td>
</tr>
<tr>
<td>RHC</td>
<td>2184</td>
<td>0.38081953</td>
</tr>
</tbody>
</table>

Summary information for training sample of size 5735
d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators),
s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
m=missing-value flag variable, p=periodic variable, w=weight

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>cat1</td>
<td>c</td>
<td></td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>cat2</td>
<td>c</td>
<td></td>
<td>6</td>
<td>4535</td>
</tr>
<tr>
<td>4</td>
<td>ca</td>
<td>c</td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>cardiohx</td>
<td>c</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>chfhx</td>
<td>c</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>dementhx</td>
<td>c</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>psychhx</td>
<td>c</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>chrpulhx</td>
<td>c</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>renalhx</td>
<td>c</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>liverhx</td>
<td>c</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>gildedhx</td>
<td>c</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>malighx</td>
<td>c</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>immunhx</td>
<td>c</td>
<td></td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
The above lists the active variables and their summary statistics.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Data Type</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>transhx</td>
<td>c</td>
<td>s</td>
<td>18.04</td>
<td>101.8</td>
</tr>
<tr>
<td>amihx</td>
<td>c</td>
<td>s</td>
<td>0.000</td>
<td>30.00</td>
</tr>
<tr>
<td>age</td>
<td>s</td>
<td>s</td>
<td>8.000</td>
<td>250.0</td>
</tr>
<tr>
<td>edu</td>
<td>s</td>
<td>s</td>
<td>0.9999E-01</td>
<td>58.20</td>
</tr>
<tr>
<td>surv2md1</td>
<td>s</td>
<td>s</td>
<td>0.000</td>
<td>0.9620</td>
</tr>
<tr>
<td>das2d3pc</td>
<td>s</td>
<td>s</td>
<td>11.00</td>
<td>33.00</td>
</tr>
<tr>
<td>aps1</td>
<td>s</td>
<td>s</td>
<td>3.000</td>
<td>147.0</td>
</tr>
<tr>
<td>scoma1</td>
<td>s</td>
<td>s</td>
<td>0.000</td>
<td>100.0</td>
</tr>
<tr>
<td>meanbp1</td>
<td>s</td>
<td>s</td>
<td>10.00</td>
<td>259.0</td>
</tr>
<tr>
<td>wblc1</td>
<td>s</td>
<td>s</td>
<td>0.000</td>
<td>192.0</td>
</tr>
<tr>
<td>hrt1</td>
<td>s</td>
<td>s</td>
<td>8.000</td>
<td>250.0</td>
</tr>
<tr>
<td>resp1</td>
<td>s</td>
<td>s</td>
<td>2.000</td>
<td>100.0</td>
</tr>
<tr>
<td>temp1</td>
<td>s</td>
<td>s</td>
<td>27.00</td>
<td>43.00</td>
</tr>
<tr>
<td>pafii1</td>
<td>s</td>
<td>s</td>
<td>11.60</td>
<td>937.5</td>
</tr>
<tr>
<td>alb1</td>
<td>s</td>
<td>s</td>
<td>0.3000</td>
<td>29.00</td>
</tr>
<tr>
<td>hemal</td>
<td>s</td>
<td>s</td>
<td>2.000</td>
<td>66.19</td>
</tr>
<tr>
<td>bil1</td>
<td>s</td>
<td>s</td>
<td>0.9999E-01</td>
<td>58.20</td>
</tr>
<tr>
<td>crea1</td>
<td>s</td>
<td>s</td>
<td>0.9999E-01</td>
<td>25.10</td>
</tr>
<tr>
<td>sod1</td>
<td>s</td>
<td>s</td>
<td>101.0</td>
<td>178.0</td>
</tr>
<tr>
<td>pot1</td>
<td>s</td>
<td>s</td>
<td>1.100</td>
<td>11.90</td>
</tr>
<tr>
<td>pac21</td>
<td>s</td>
<td>s</td>
<td>1.000</td>
<td>156.0</td>
</tr>
<tr>
<td>ph1</td>
<td>s</td>
<td>s</td>
<td>6.579</td>
<td>7.770</td>
</tr>
<tr>
<td>swang1</td>
<td>d</td>
<td>s</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>wtkilo1</td>
<td>s</td>
<td>s</td>
<td>19.50</td>
<td>244.0</td>
</tr>
<tr>
<td>dnri</td>
<td>c</td>
<td>s</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ninsclas</td>
<td>c</td>
<td>s</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>resp</td>
<td>c</td>
<td>s</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>card</td>
<td>c</td>
<td>s</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>neuro</td>
<td>c</td>
<td>s</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>gastr</td>
<td>c</td>
<td>s</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>renal</td>
<td>c</td>
<td>s</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>meta</td>
<td>c</td>
<td>s</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>hema</td>
<td>c</td>
<td>s</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>seps</td>
<td>c</td>
<td>s</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>trauma</td>
<td>c</td>
<td>s</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ortho</td>
<td>c</td>
<td>s</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>adddp</td>
<td>s</td>
<td>s</td>
<td>0.000</td>
<td>7.000</td>
</tr>
<tr>
<td>urin1</td>
<td>s</td>
<td>s</td>
<td>0.000</td>
<td>9000.</td>
</tr>
<tr>
<td>race</td>
<td>c</td>
<td>s</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>income</td>
<td>c</td>
<td>s</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

The total #cases w/ #missing

<table>
<thead>
<tr>
<th>Total</th>
<th>#cases w/</th>
<th>#missing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#cases miss. D ord. vals</td>
<td>#X-var</td>
</tr>
<tr>
<td>5735</td>
<td>0</td>
<td>5157</td>
</tr>
</tbody>
</table>
Number of cases used for training: 5735
Number of split variables: 53
Number of cases excluded due to 0 W or missing D variable: 0

Constant fitted to cases with missing values in regressor variables
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE’s for pruned tree: 0.2500

Node predictions are made by majority rule.
Class priors estimated by sample proportions.
Unit misclassification costs

Top-ranked variables and 1-df chi-squared values at root node
1. 0.3346E+03 cat1
2. 0.2728E+03 aps1
3. 0.2430E+03 crea1
... 50. 0.1052E+01 meta
51. 0.6357E+00 race

Size and CV mean cost and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Nodes</th>
<th>Mean Cost</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median Cost</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68</td>
<td>3.236E-01</td>
<td>6.178E-03</td>
<td>3.960E-03</td>
<td>3.284E-01</td>
<td>6.780E-03</td>
</tr>
<tr>
<td>2</td>
<td>67</td>
<td>3.236E-01</td>
<td>6.178E-03</td>
<td>3.960E-03</td>
<td>3.284E-01</td>
<td>6.780E-03</td>
</tr>
<tr>
<td>3</td>
<td>66</td>
<td>3.236E-01</td>
<td>6.178E-03</td>
<td>3.960E-03</td>
<td>3.284E-01</td>
<td>6.780E-03</td>
</tr>
<tr>
<td>4</td>
<td>65</td>
<td>3.236E-01</td>
<td>6.178E-03</td>
<td>3.960E-03</td>
<td>3.284E-01</td>
<td>6.780E-03</td>
</tr>
</tbody>
</table>
... 37  18  3.180E-01  6.150E-03  2.945E-03  3.217E-01  3.907E-03
38+  12  3.198E-01  6.159E-03  3.064E-03  3.182E-01  3.105E-03
39** 10  3.180E-01  6.150E-03  2.127E-03  3.188E-01  3.098E-03
40   8   3.219E-01  6.169E-03  3.105E-03  3.217E-01  5.293E-03
41   6   3.240E-01  6.180E-03  3.474E-03  3.249E-01  6.673E-03
42   5   3.228E-01  6.174E-03  3.471E-03  3.249E-01  5.539E-03
43   3   3.325E-01  6.221E-03  3.356E-03  3.365E-01  6.220E-03
44   2   3.751E-01  6.393E-03  4.245E-03  3.801E-01  3.186E-03
45   1   3.808E-01  6.412E-03  2.782E-04  3.805E-01  4.832E-04

Above shows that the largest tree has 68 terminal nodes.
0-SE tree based on mean is marked with * and has 10 terminal nodes
0-SE tree based on median is marked with + and has 12 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
** tree same as ++ tree
** tree same as -- tree
++ tree same as -- tree
* tree same as ** tree
* tree same as ++ tree
* tree same as -- tree

Pruned tree has 10 terminal nodes and is marked by two asterisks.

Following tree is based on mean CV with naive SE estimate (**)  

Structure of final tree. Each terminal node is marked with a T.

<table>
<thead>
<tr>
<th>Node</th>
<th>Total cases</th>
<th>Train cases</th>
<th>Predicted class</th>
<th>Node</th>
<th>Split variable</th>
<th>Interacting variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5735</td>
<td>5735</td>
<td>NoRHC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1683</td>
<td>1683</td>
<td>RHC</td>
<td>3.808E-01</td>
<td>cat1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1117</td>
<td>1117</td>
<td>RHC</td>
<td>3.796E-01</td>
<td>pafi1</td>
<td></td>
</tr>
<tr>
<td>8T</td>
<td>655</td>
<td>655</td>
<td>RHC</td>
<td>3.038E-01</td>
<td>resp1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>462</td>
<td>462</td>
<td>RHC</td>
<td>4.870E-01</td>
<td>ninsclas</td>
<td></td>
</tr>
<tr>
<td>18T</td>
<td>244</td>
<td>244</td>
<td>RHC</td>
<td>3.730E-01</td>
<td>bili1</td>
<td></td>
</tr>
<tr>
<td>19T</td>
<td>218</td>
<td>218</td>
<td>NoRHC</td>
<td>3.853E-01</td>
<td>card</td>
<td></td>
</tr>
<tr>
<td>5T</td>
<td>566</td>
<td>566</td>
<td>NoRHC</td>
<td>3.816E-01</td>
<td>alb1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>4052</td>
<td>4052</td>
<td>NoRHC</td>
<td>3.147E-01</td>
<td>pafi1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1292</td>
<td>1292</td>
<td>NoRHC</td>
<td>4.837E-01</td>
<td>resp</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>581</td>
<td>581</td>
<td>RHC</td>
<td>4.200E-01</td>
<td>dnr1</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>515</td>
<td>515</td>
<td>RHC</td>
<td>3.903E-01</td>
<td>cat1</td>
<td></td>
</tr>
<tr>
<td>48T</td>
<td>438</td>
<td>438</td>
<td>RHC</td>
<td>3.447E-01</td>
<td>meanbp1</td>
<td></td>
</tr>
<tr>
<td>49T</td>
<td>77</td>
<td>77</td>
<td>NoRHC</td>
<td>3.506E-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25T</td>
<td>66</td>
<td>66</td>
<td>NoRHC</td>
<td>3.485E-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>711</td>
<td>711</td>
<td>NoRHC</td>
<td>4.051E-01</td>
<td>seps</td>
<td></td>
</tr>
<tr>
<td>26T</td>
<td>110</td>
<td>110</td>
<td>RHC</td>
<td>3.636E-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27T</td>
<td>601</td>
<td>601</td>
<td>NoRHC</td>
<td>3.627E-01</td>
<td>adld3p</td>
<td></td>
</tr>
<tr>
<td>7T</td>
<td>2760</td>
<td>2760</td>
<td>NoRHC</td>
<td>2.355E-01</td>
<td>apsi</td>
<td></td>
</tr>
</tbody>
</table>

Above gives the number of observations in each node (terminal node marked with a T), its predicted class, and the split variable.

Number of terminal nodes of final tree: 10
Total number of nodes of final tree: 19
Second best split variable (based on curvature test) at root node is apsi
If cat1 is omitted, apsi will be chosen to split the root node.

Classification tree:
For categorical variable splits, values not in training data go to the right
4.1 Univariate splits

Node 1: cat1 = "CHF", "MOSF w/Sepsis"
   Node 2: meanbp1 <= 68.500000 or NA
   Node 4: pafi1 <= 266.15625
      Node 8: RHC
   Node 4: pafi1 > 266.15625 or NA
      Node 9: ninsclas = "No insurance", "Private", "Private & Medicare"
         Node 18: RHC
      Node 9: ninsclas /= "No insurance", "Private", "Private & Medicare"
         Node 19: NoRHC
   Node 2: meanbp1 > 68.500000
   Node 5: NoRHC
Node 1: cat1 /= "CHF", "MOSF w/Sepsis"
   Node 3: pafi1 <= 142.35938
   Node 6: resp = "No"
      Node 12: dnr1 = "No"
         Node 24: cat1 = "ARF", "Lung Cancer", "MOSF w/Malignancy"
            Node 48: RHC
         Node 24: cat1 /= "ARF", "Lung Cancer", "MOSF w/Malignancy"
            Node 49: NoRHC
      Node 12: dnr1 /= "No"
      Node 25: NoRHC
   Node 6: resp /= "No"
   Node 13: seps = "Yes"
      Node 26: RHC
   Node 13: seps /= "Yes"
      Node 27: NoRHC
   Node 3: pafi1 > 142.35938 or NA
   Node 7: NoRHC

***************************************************************************
Predictor means below are means of cases with no missing values.

Node 1: Intermediate node
   A case goes into Node 2 if cat1 = "CHF", "MOSF w/Sepsis"
   cat1 mode = "ARF"
Class            Number  Posterior
NoRHC            3551  0.6192E+00
RHC              2184  0.3808E+00
Number of training cases misclassified = 2184
Predicted class is NoRHC

Node 2: Intermediate node
   A case goes into Node 4 if meanbp1 <= 68.500000 or NA
   meanbp1 mean = 72.674985
Class            Number  Posterior
NoRHC            774   0.4599E+00

Wei-Yin Loh

27

GUIDE manual
4.1 Univariate splits

RHC 909 0.5401E+00
Number of training cases misclassified = 774
Predicted class is RHC

Node 4: Intermediate node
A case goes into Node 8 if pafi1 <= 266.15625
pafi1 mean = 241.37331

Class Number Posterior
NoRHC 424 0.3796E+00
RHC 693 0.6204E+00
Number of training cases misclassified = 424
Predicted class is RHC

Node 8: Terminal node

Class Number Posterior
NoRHC 199 0.3038E+00
RHC 456 0.6962E+00
Number of training cases misclassified = 199
Predicted class is RHC

Node 9: Intermediate node
A case goes into Node 18 if ninsclas = "No insurance", "Private", "Private & Medicare"
ninsclas mode = "Private"

Class Number Posterior
NoRHC 225 0.4870E+00
RHC 237 0.5130E+00
Number of training cases misclassified = 225
Predicted class is RHC

Node 18: Terminal node

Class Number Posterior
NoRHC 91 0.3730E+00
RHC 153 0.6270E+00
Number of training cases misclassified = 91
Predicted class is RHC

Node 19: Terminal node

Class Number Posterior
NoRHC 134 0.6147E+00
RHC 84 0.3853E+00
Number of training cases misclassified = 84
Predicted class is NoRHC

Node 5: Terminal node

Class Number Posterior
4.1 Univariate splits

<table>
<thead>
<tr>
<th>Class</th>
<th>Number</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoRHC</td>
<td>350</td>
<td>0.6184E+00</td>
</tr>
<tr>
<td>RHC</td>
<td>216</td>
<td>0.3816E+00</td>
</tr>
</tbody>
</table>

Number of training cases misclassified = 216
Predicted class is NoRHC

----------------------------

Node 3: Intermediate node
A case goes into Node 6 if pafi1 <= 142.35938
pafi1 mean = 211.08630

<table>
<thead>
<tr>
<th>Class</th>
<th>Number</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoRHC</td>
<td>2777</td>
<td>0.6853E+00</td>
</tr>
<tr>
<td>RHC</td>
<td>1275</td>
<td>0.3147E+00</td>
</tr>
</tbody>
</table>

Number of training cases misclassified = 1275
Predicted class is NoRHC

----------------------------

Node 6: Intermediate node
A case goes into Node 12 if resp = "No"
resp mode = "Yes"

<table>
<thead>
<tr>
<th>Class</th>
<th>Number</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoRHC</td>
<td>667</td>
<td>0.5163E+00</td>
</tr>
<tr>
<td>RHC</td>
<td>625</td>
<td>0.4837E+00</td>
</tr>
</tbody>
</table>

Number of training cases misclassified = 625
Predicted class is NoRHC

----------------------------

Node 12: Intermediate node
A case goes into Node 24 if dnr1 = "No"
dnr1 mode = "No"

<table>
<thead>
<tr>
<th>Class</th>
<th>Number</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoRHC</td>
<td>244</td>
<td>0.4200E+00</td>
</tr>
<tr>
<td>RHC</td>
<td>337</td>
<td>0.5800E+00</td>
</tr>
</tbody>
</table>

Number of training cases misclassified = 244
Predicted class is RHC

----------------------------

Node 24: Intermediate node
A case goes into Node 48 if cat1 = "ARF", "Lung Cancer", "MOSF w/Malignancy"
cat1 mode = "ARF"

<table>
<thead>
<tr>
<th>Class</th>
<th>Number</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoRHC</td>
<td>201</td>
<td>0.3903E+00</td>
</tr>
<tr>
<td>RHC</td>
<td>314</td>
<td>0.6097E+00</td>
</tr>
</tbody>
</table>

Number of training cases misclassified = 201
Predicted class is RHC

----------------------------

Node 48: Terminal node
class Number | Posterior |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NoRHC</td>
<td>151</td>
</tr>
<tr>
<td>RHC</td>
<td>287</td>
</tr>
</tbody>
</table>

Number of training cases misclassified = 151
Predicted class is RHC

Node 49: Terminal node
Class        Number  Posterior
NoRHC        50       0.6494E+00
RHC          27       0.3506E+00
Number of training cases misclassified = 27
Predicted class is NoRHC

Node 25: Terminal node
Class        Number  Posterior
NoRHC        43       0.6515E+00
RHC          23       0.3485E+00
Number of training cases misclassified = 23
Predicted class is NoRHC

Node 13: Intermediate node
A case goes into Node 26 if seps = "Yes"
seps mode = "No"
Class        Number  Posterior
NoRHC        423      0.5949E+00
RHC          288      0.4051E+00
Number of training cases misclassified = 288
Predicted class is NoRHC

Node 26: Terminal node
Class        Number  Posterior
NoRHC        40       0.3636E+00
RHC          70       0.6364E+00
Number of training cases misclassified = 40
Predicted class is RHC

Node 27: Terminal node
Class        Number  Posterior
NoRHC        383      0.6373E+00
RHC          218      0.3627E+00
Number of training cases misclassified = 218
Predicted class is NoRHC

Node 7: Terminal node
Class        Number  Posterior
NoRHC        2110     0.7645E+00
RHC          650      0.2355E+00
Number of training cases misclassified = 650
Predicted class is NoRHC
4.1 Univariate splits

Classification matrix for training sample:

<table>
<thead>
<tr>
<th>Predicted</th>
<th>True class</th>
</tr>
</thead>
<tbody>
<tr>
<td>class</td>
<td>NoRHC</td>
</tr>
<tr>
<td>NoRHC</td>
<td>3070</td>
</tr>
<tr>
<td>RHC</td>
<td>481</td>
</tr>
<tr>
<td>Total</td>
<td>3551</td>
</tr>
</tbody>
</table>

Number of cases used for tree construction: 5735
Number misclassified: 1699
Resubstitution estimate of mean misclassification cost: 0.29625109

Resubstitution estimate = (number misclassified)/(number of cases).

Observed and fitted values are stored in classfit.txt
LaTeX code for tree is in class.tex
R code is stored in classpred.r
Elapsed time in seconds: 14.489

Figure 1 shows the \LaTeX tree. Symbol “≤∗” in the split at node 2, “meanbp1 ≤∗ 68.50”, means that observations with missing values in the variable go left. If missing values go right, as in node 3, there is no asterisk beside the inequality sign. The tree diagram can be viewed and saved as pdf by following the directions on page 11.

4.1.4 Contents of classfit.txt

Below are the first few lines of the file classfit.txt.

| train | node | observed | predicted | "P(NoRHC)" | "P(RHC)"
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>27</td>
<td>&quot;NoRHC&quot;</td>
<td>&quot;NoRHC&quot;</td>
<td>0.63727E+00</td>
<td>0.36273E+00</td>
</tr>
<tr>
<td>y</td>
<td>8</td>
<td>&quot;RHC&quot;</td>
<td>&quot;RHC&quot;</td>
<td>0.30382E+00</td>
<td>0.69618E+00</td>
</tr>
<tr>
<td>y</td>
<td>7</td>
<td>&quot;RHC&quot;</td>
<td>&quot;NoRHC&quot;</td>
<td>0.76449E+00</td>
<td>0.23551E+00</td>
</tr>
<tr>
<td>y</td>
<td>7</td>
<td>&quot;NoRHC&quot;</td>
<td>&quot;NoRHC&quot;</td>
<td>0.76449E+00</td>
<td>0.23551E+00</td>
</tr>
<tr>
<td>y</td>
<td>19</td>
<td>&quot;RHC&quot;</td>
<td>&quot;NoRHC&quot;</td>
<td>0.61468E+00</td>
<td>0.38532E+00</td>
</tr>
</tbody>
</table>

The row in this file match those in the data file. The meanings of the columns are:

**train**: equals “y” (for “yes”) if the observation was used in model construction; otherwise “n” (for “no”). All the values in this example are “y” because every observation is used. Two typical situations where this value is n are (i) if its d variable value is missing and (ii) if there is a weight variable in the data that takes value 0 for the observation.
Figure 1: GUIDE v.41.0 0.250-SE classification tree for predicting `swang1` using estimated priors and unit misclassification costs. At each split, an observation goes to the left branch if and only if the condition is satisfied. Symbol ‘≤’ stands for ‘≤’ or missing’. $S_1 = \{\text{CHF, MOSF w/Sepsis}\}$. $S_2 = \{\text{No insurance, Private, Private & Medicare}\}$. $S_3 = \{\text{ARF, Lung Cancer, MOSF w/Malignancy}\}$. Predicted classes and sample sizes (in *italics*) printed below terminal nodes; class sample proportion for `swang1 = RHC` beside nodes. Second best split variable at root node is `aps1`. 
node: label of the terminal node the observation belongs to. For example, the first observation landed in node 27.

observed: value of the d variable for this observation in the data file.

predicted: predicted value of the d variable for this observation.

P(NoRHC): estimated posterior probability that the observation is in class “NoRHC”.

P(RHC): estimated posterior probability that the observation is in class “RHC”.

The posterior probabilities are calculated as follows. Let J be the number of classes, \( N_j \) be the number of class \( j \) observations in the whole sample and \( N = \sum_j N_j \). Let \( \pi_j \) be the (estimated or specified) prior probability of class \( j \). Let \( n_j(t) \) be the number of class \( j \) training samples in node \( t \). The posterior probability of class \( j \) in \( t \) is \( p_j(t) = \pi_j n_j(t) N_j^{-1} / \sum_i \pi_i n_i(t) N_i^{-1} \). If \( \min_j p_j(t) = 0 \), the posterior probability is redefined to be \( (N \pi_j(t) + \pi_j) / (N + 1) \); this ensures that no probability is zero if all \( \pi_j \) are positive.

### 4.1.5 Contents of classpred.r

The file `classpred.r` gives an R function for computing the predicted class and posterior probabilities.

```r
predicted <- function(){
  catvalues <- c("CHF","MOSF w/Sepsis")
  if(cat1 %in% catvalues){
    if(is.na(meanbp1) | meanbp1 <= 68.5000000000 ){
      if(!is.na(pafi1) & pafi1 <= 266.156250000 ){
        nodeid <- 8
        predclass <- "RHC"
        posterior <- c( 0.30382E+00, 0.69618E+00)
      } else {
        catvalues <- c("No insurance","Private","Private & Medicare")
        if(ninsclas %in% catvalues){
          nodeid <- 18
          predclass <- "RHC"
          posterior <- c( 0.37295E+00, 0.62705E+00)
        } else {
          nodeid <- 19
          predclass <- "NoRHC"
          posterior <- c( 0.61468E+00, 0.38532E+00)
        }
      }
    } else {
      catvalues <- c("No insurance","Private","Private & Medicare")
      if(ninsclas %in% catvalues){
        nodeid <- -18
        predclass <- "RHC"
        posterior <- c( 0.30382E+00, 0.69618E+00)
      } else {
        nodeid <- -19
        predclass <- "NoRHC"
        posterior <- c( 0.61468E+00, 0.38532E+00)
      }
    }
  }
}
```
nodeid <- 5
predclass <- "NoRHC"
predclass <- "NoRHC"
posterior <- c(0.61837E+00, 0.38163E+00)
}
} else {
  if(!is.na(pafi1) & pafi1 <= 142.359375000 ){
    catvalues <- c("No")
    if(resp %in% catvalues){
      catvalues <- c("No")
      if(dnr1 %in% catvalues){
        catvalues <- c("ARF","Lung Cancer","MOSF w/Malignancy")
        if(cat1 %in% catvalues){
          nodeid <- 48
          predclass <- "RHC"
          posterior <- c(0.34475E+00, 0.65525E+00)
        } else {
          nodeid <- 49
          predclass <- "NoRHC"
          posterior <- c(0.64935E+00, 0.35065E+00)
        }
      } else {
        nodeid <- 25
        predclass <- "NoRHC"
        posterior <- c(0.65152E+00, 0.34848E+00)
      }
    } else {
      nodeid <- 26
      predclass <- "RHC"
      posterior <- c(0.36364E+00, 0.63636E+00)
    }
  } else {
    catvalues <- c("Yes")
    if(seps %in% catvalues){
      nodeid <- 26
      predclass <- "RHC"
      posterior <- c(0.36364E+00, 0.63636E+00)
    } else {
      nodeid <- 27
      predclass <- "NoRHC"
      posterior <- c(0.63727E+00, 0.36273E+00)
    }
  } else {
    nodeid <- 7
    predclass <- "NoRHC"
    posterior <- c(0.76449E+00, 0.23551E+00)
  }
}
return(c(nodeid,predclass,posterior))

## end of function
4.2 Linear splits

The classification tree in Figure 1 can sometimes be reduced in size if we employ two ordinal variables to split each node. This can be done by selecting a non-default option.

4.2.1 Input file generation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: linearin.txt
   Input 1 for model fitting, 2 for importance or DIF scoring, 3 for data conversion ([1:3], <cr>=1): 1
   Name of batch output file: linearout.txt
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
4.2 Linear splits

Input your choice ([1:3], <cr>=1):
Input 1 for default options, 2 otherwise ([1:2], <cr>=1): 2
Input 1 for simple, 2 for nearest-neighbor, 3 for kernel method ([1:3], <cr>=1):
Input 0 for linear, interaction and univariate splits (in this order),
1 for univariate, linear and interaction splits (in this order),
2 to skip linear splits,
3 to skip linear and interaction splits:
Input your choice ([0:3], <cr>=1): 0
Input 0 to specify tree with fixed no. of nodes, 1 to prune by CV, 2
by test sample, 3 for no pruning ([0:3], <cr>=1):
Input name of data description file (max 100 characters);
enclose with matching quotes if it has spaces: rhcsc1.txt
Reading data description file ...
Training sample file: rhcdata.txt
Missing value code: NA
Records in data file start on line 2
23 N variables changed to S
D variable is swang1
Reading data file ...
Number of records in data file: 5735
Length of longest entry in data file: 19
Checking for missing values ...
Finished checking
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Recoding D values to integers
Finished recoding
Number of classes: 2
Assigning integer codes to values of 30 categorical variables
Finished assigning codes to 10 categorical variables
Finished assigning codes to 20 categorical variables
Finished assigning codes to 30 categorical variables
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...
Finished processing 5000 of 5735 observations
Data checks complete
Creating missing value indicators ...
Rereading data ...

Class       #Cases    Proportion
NoRHC       3551     0.61918047
RHC         2184     0.38081953
Total  #cases w/    #missing
      #cases      miss.  D   ord.   vals   #X-var  #N-var  #F-var  #S-var
      5735        0     5157     10       0       0       0       23

Wei-Yin Loh 36 GUIDE manual
4.2 Linear splits

<table>
<thead>
<tr>
<th>P-var</th>
<th>M-var</th>
<th>B-var</th>
<th>C-var</th>
<th>I-var</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>30</td>
<td>0</td>
</tr>
</tbody>
</table>

Number of cases used for training: 5735
Number of split variables: 53
Number of cases excluded due to 0 W or missing D variable: 0
Finished reading data file
Default number of cross-validations: 10
Input 1 to accept the default, 2 to change it ([1:2], <cr>=1):
Best tree may be chosen based on mean or median CV estimate
Input 1 for mean-based, 2 for median-based ([1:2], <cr>=1):
Input number of SEs for pruning ([0.00:1000.00], <cr>=0.25):
Choose 1 for estimated priors, 2 for equal priors, 3 to input priors from a file
Input 1, 2, or 3 ([1:3], <cr>=1):
Choose 1 for unit misclassification costs, 2 to input costs from a file
Input 1 or 2 ([1:2], <cr>=1):
Choose a split point selection method for numerical variables:
Choose 1 to use faster method based on sample quantiles
Choose 2 to use exhaustive search
Input 1 or 2 ([1:2], <cr>=2):
Default max. number of split levels: 15
Input 1 to accept this value, 2 to change it ([1:2], <cr>=1):
Default minimum node sample size is 57
Input 1 to use the default value, 2 to change it ([1:2], <cr>=1):
Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):
Input file name to store LaTeX code (use .tex as suffix): linear.tex
Input 1 to color terminal nodes, 2 otherwise ([1:2], <cr>=1):
Choose amount of detail in nodes of LaTeX tree diagram:
Input 0 for #errors, 1 for sample sizes, 2 for sample proportions, 3 for posterior probs,
4 for nothing
Input your choice ([0:4], <cr>=2):
You can store the variables and/or values used to split and fit in a file
Choose 1 to skip this step, 2 to store split variables and their values
Input your choice ([1:2], <cr>=1):
Input 2 to save fitted values and node IDs, 1 otherwise ([1:2], <cr>=2):
Input name of file to store node ID and fitted value of each case: linearfit.txt
Input 2 to write R function for predicting new cases, 1 otherwise ([1:2], <cr>=2):
Input file name: linearpred.r
Input rank of top variable to split root node ([1:53], <cr>=1):
Input file is created!
Run GUIDE with the command: guide < linearin.txt
Press ENTER or RETURN to quit

4.2.2 Contents of linearin.txt

GUIDE (do not edit this file unless you know what you are doing)

Wei-Yin Loh  37  GUIDE manual
4.2 Linear splits

4.2.3 Contents of linearout.txt

Classification tree
Pruning by cross-validation
Data description file: rhcdsc1.txt
Training sample file: rhcdata.txt
Missing value code: NA
Records in data file start on line 2
23 N variables changed to S
D variable is swang1
Number of records in data file: 5735
Length of longest entry in data file: 19
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Number of classes: 2
Training sample class proportions of D variable swang1:
### 4.2 Linear splits

**CLASSIFICATION: RHC DATA**

<table>
<thead>
<tr>
<th>Class</th>
<th>#Cases</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoRHC</td>
<td>3551</td>
<td>0.61918047</td>
</tr>
<tr>
<td>RHC</td>
<td>2184</td>
<td>0.38081953</td>
</tr>
</tbody>
</table>

Summary information for training sample of size 5735

d=dependent, b=split and fit cat variable using indicator variables,  
c=split-only categorical, i=fit-only categorical (via indicators),  
s=split-only numerical, n=split and fit numerical, f=fit-only numerical,  
m=missing-value flag variable, p=periodic variable, w=weight

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>cat1</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>cat2</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>ca</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>cardiohx</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>chfhx</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>dementhx</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>psychhx</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>chrpulhx</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>renalhx</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>liverhx</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>gibledhx</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>malighx</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>immunhx</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>transhx</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>amihx</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>age</td>
<td>s</td>
<td>18.04</td>
<td>101.8</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>sex</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>edu</td>
<td>s</td>
<td>0.000</td>
<td>30.00</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>surv2md1</td>
<td>s</td>
<td>0.000</td>
<td>0.9620</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>das2d3pc</td>
<td>s</td>
<td>11.00</td>
<td>33.00</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>aps1</td>
<td>s</td>
<td>3.000</td>
<td>147.0</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>scoma1</td>
<td>s</td>
<td>0.000</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>meanbp1</td>
<td>s</td>
<td>10.00</td>
<td>259.0</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>wblc1</td>
<td>s</td>
<td>0.000</td>
<td>192.0</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>hrt1</td>
<td>s</td>
<td>8.000</td>
<td>250.0</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>resp1</td>
<td>s</td>
<td>2.000</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>temp1</td>
<td>s</td>
<td>27.00</td>
<td>43.00</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>pafi1</td>
<td>s</td>
<td>11.60</td>
<td>937.5</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>alb1</td>
<td>s</td>
<td>0.3000</td>
<td>29.00</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>hema1</td>
<td>s</td>
<td>2.000</td>
<td>66.19</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>bili1</td>
<td>s</td>
<td>0.9999E-01</td>
<td>58.20</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>creat</td>
<td>s</td>
<td>0.9999E-01</td>
<td>25.10</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>sod1</td>
<td>s</td>
<td>101.0</td>
<td>178.0</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>pot1</td>
<td>s</td>
<td>1.100</td>
<td>11.90</td>
<td></td>
</tr>
</tbody>
</table>

Wei-Yin Loh 39 GUIDE manual
4.2 Linear splits

43 paco21 s 1.000 156.0
44 ph1 s 6.579 7.770
45 swang1 d 2
46 wtkilo1 s 19.50 244.0 515
47 dnr1 c 2
48 ninsclas c 6
49 resp c 2
50 card c 2
51 neuro c 2
52 gastr c 2
53 renal c 2
54 meta c 2
55 hema c 2
56 seps c 2
57 trauma c 2
58 ortho c 2
59 adld3p s 0.000 7.000 4296
60 urin1 s 0.000 9000 3028
61 race c 3
62 income c 4

Total #cases w/ #missing
#cases miss. D ord. vals #X-var #N-var #F-var #S-var
5735 0 5157 10 0 0 23

#P-var #M-var #B-var #C-var #I-var
0 0 0 30 0

Number of cases used for training: 5735
Number of split variables: 53
Number of cases excluded due to 0 W or missing D variable: 0

Constant fitted to cases with missing values in regressor variables
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE's for pruned tree: 0.2500

Simple node models
Estimated priors
Unit misclassification costs
Linear split highest priority
Interaction and linear splits 2nd and 3rd priorities
Split values for N and S variables based on exhaustive search
Maximum number of split levels: 15
Minimum node sample size: 57
Top-ranked variables and 1-df chi-squared values at root node
1 0.3346E+03 cat1
2 0.2728E+03 aps1
4.2 Linear splits

<table>
<thead>
<tr>
<th>Node</th>
<th>Total</th>
<th>Train</th>
<th>Predicted</th>
<th>Node Split</th>
<th>Interacting</th>
<th>cost</th>
<th>variables</th>
<th>variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5735</td>
<td>5735</td>
<td>NoRHC</td>
<td>3.808E-01</td>
<td>cat1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1683</td>
<td>1683</td>
<td>RHC</td>
<td>4.599E-01</td>
<td>meanbp1 +pafi1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1174</td>
<td>1174</td>
<td>RHC</td>
<td>3.705E-01</td>
<td>resl1 +surv2md1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8T</td>
<td>229</td>
<td>229</td>
<td>RHC</td>
<td>1.790E-01</td>
<td>sod1 :wtkilo1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>945</td>
<td>945</td>
<td>RHC</td>
<td>4.169E-01</td>
<td>ninsclas</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18T</td>
<td>321</td>
<td>321</td>
<td>RHC</td>
<td>3.084E-01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>624</td>
<td>624</td>
<td>RHC</td>
<td>4.728E-01</td>
<td>drn1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>554</td>
<td>554</td>
<td>RHC</td>
<td>4.495E-01</td>
<td>adld3p +edu</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>76T</td>
<td>479</td>
<td>479</td>
<td>RHC</td>
<td>4.071E-01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>77T</td>
<td>75</td>
<td>75</td>
<td>NoRHC</td>
<td>2.800E-01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.2 Linear splits

Number of terminal nodes of final tree: 16
Total number of nodes of final tree: 31
Second best split variable (based on curvature test) at root node is aps1

Classification tree:
For categorical variable splits, values not in training data go to the right

Node 1: cat1 = "CHF", "MOSF w/Sepsis"
  Node 2: 0.24316737 * pafi1 + meanbp1 <= 153.28329 or NA
    Node 4: 48.127695 * surv2md1 + resp1 <= 43.437797 or NA
      Node 8: RHC
    Node 4: 48.127695 * surv2md1 + resp1 > 43.437797
      Node 9: ninsclas = "No insurance", "Private"
      Node 18: RHC
    Node 9: ninsclas /= "No insurance", "Private"
      Node 19: dnr1 = "No"
        Node 38: -23.826398 * edu + adld3p <= -282.91678 or NA
          Node 76: RHC
        Node 38: -23.826398 * edu + adld3p > -282.91678
          Node 77: NoRHC
      Node 19: dnr1 /= "No"
      Node 39: NoRHC
  Node 2: 0.24316737 * pafi1 + meanbp1 > 153.28329
    Node 5: NoRHC
4.2 Linear splits

Node 1: cat1 /= "CHF", "MOSF w/Sepsis"
Node 3: 11.508773 * adld3p + pafi1 <= 149.35252 or NA
  Node 6: -1.3120163 * hema1 + aps1 <= 0.84337055
    Node 12: NoRHC
  Node 6: -1.3120163 * hema1 + aps1 > 0.84337055 or NA
  Node 13: 4.0975611 * resp1 + pafi1 <= 207.99333
    Node 26: RHC
  Node 13: 4.0975611 * resp1 + pafi1 > 207.99333 or NA
  Node 27: -23.161068 * adld3p + aps1 <= 66.838932
    Node 54: NoRHC
  Node 27: -23.161068 * adld3p + aps1 > 66.838932 or NA
  Node 55: 1.0116045 * wtkilo1 + aps1 <= 121.69374 or NA
    Node 110: NoRHC
  Node 55: 1.0116045 * wtkilo1 + aps1 > 121.69374
  Node 111: 0.35358803 * pafi1 + meanbp1 <= 134.65949 or NA
    Node 222: -0.42185873 * wtkilo1 + paco21 <= -7.0243280
      Node 444: RHC
    Node 222: -0.42185873 * wtkilo1 + paco21 > -7.0243280 or NA
      Node 445: scoma1 <= 4.5000000
        Node 890: 5.8542561 * pot1 + bili1 <= 25.404949
          Node 1780: RHC
        Node 890: 5.8542561 * pot1 + bili1 > 25.404949 or NA
          Node 1781: NoRHC
        Node 445: scoma1 > 4.5000000 or NA
          Node 891: NoRHC
    Node 111: 0.35358803 * pafi1 + meanbp1 > 134.65949
    Node 223: NoRHC
Node 3: 11.508773 * adld3p + pafi1 > 149.35252
Node 7: NoRHC

******************************************************************************
Predictor means below are means of cases with no missing values.

Node 1: Intermediate node
A case goes into Node 2 if cat1 = "CHF", "MOSF w/Sepsis"
cat1 mode = "ARF"
Class Number Posterior
NoRHC 3551 0.6192E+00
RHC 2184 0.3808E+00
Number of training cases misclassified = 2184
Predicted class is NoRHC

Node 2: Intermediate node
A case goes into Node 4 if 0.24316737 * pafi1 + meanbp1 <= 153.28329
Linear combination mean = 133.36641
Class Number Posterior
4.2 Linear splits

NoRHC    774 0.4599E+00
RHC      909 0.5401E+00
Number of training cases misclassified = 774
Predicted class is RHC

Node 891: Terminal node
Class     Number  Posterior
NoRHC    137 0.5524E+00
RHC      111 0.4476E+00
Number of training cases misclassified = 111
Predicted class is NoRHC

Node 223: Terminal node
Class     Number  Posterior
NoRHC    370 0.6549E+00
RHC      195 0.3451E+00
Number of training cases misclassified = 195
Predicted class is NoRHC

Node 7: Terminal node
Class     Number  Posterior
NoRHC    621 0.8601E+00
RHC      101 0.1399E+00
Number of training cases misclassified = 101
Predicted class is NoRHC

Classification matrix for training sample:
Predicted  True class
       class  NoRHC  RHC
       NoRHC  3027  1040
       RHC    524  1144
Total     3551  2184

Number of cases used for tree construction: 5735
Number misclassified: 1564
Resubstitution estimate of mean misclassification cost: 0.27271142

Observed and fitted values are stored in linearfit.txt
LaTeX code for tree is in linear.tex
R code is stored in linearpred.r

The $\LaTeX$ tree is shown in Figure 2, where each node that is split on a pair of
4.2 Linear splits

Figure 2: GUIDE v.41.0 0.250-SE classification tree for predicting \( \text{swang1} \) using linear split priority, estimated priors and unit misclassification costs. At each split, an observation goes to the left branch if and only if the condition is satisfied. An asterisk at a bivariate split indicates that missing values in either variable go to the left node. \( S_1 = \{ \text{CHF, MOSF w/Sepsis} \} \). \( S_2 = \{ \text{No insurance, Private} \} \). Circles with dashed lines are nodes with no significant split variables. Intermediate nodes in gray indicate linear splits. Predicted classes and sample sizes (in italics) printed below terminal nodes; class sample proportion for \( \text{swang1} = \text{RHC} \) beside nodes. Second best split variable at root node is \( \text{aps1} \).

Wei-Yin Loh 45 GUIDE manual
ordi nal variables is painted gray. For example, node 2 is split on variables meanbp1 and pafi1, with observations going left if and only if

$$0.24316737 \times pafi1 + \text{meanbp1} \leq 153.28329.$$ 

The asterisk beside the node indicates that observations with missing values in either of the split variables go left. A plot of the data in this node is shown in Figure 3. The R code for making the plot is below. It reads linearfit.txt to extract the observations in the node.

### 4.2.4 R code for plot

```r
z0 <- read.table("rhcdata.txt",header=TRUE)
z1 <- read.table("linearfit.txt",header=TRUE)
gp <- z1$node == 5 | z1$node == 8 | z1$node == 18 | z1$node == 39 |
    z1$node == 76 | z1$node == 77
x <- z0$pafi1[gp]
y <- z0$meanbp1[gp]
leg.txt <- c("NoRHC","RHC")
leg.col <- c("red","blue")
```

Figure 3: Plot of meanbp1 vs pafi1 for data and split in node 2 of tree in Figure 2
4.3 Kernel discriminant models

Another way to reduce the size of a classification tree is to fit a kernel discriminant model in each node.

4.3.1 Input file generation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   1
   Name of batch input file: ker2.in
   ker2.in
   File ker2.in exists
   Input 1 to overwrite it, 2 to choose another name ([1:2], <cr>=1):

   Input 1 for model fitting, 2 for importance or DIF scoring, 3 for data conversion ([1:3], <cr>=1):

   Name of batch output file: ker2.out
   ker2.out
   File ker2.out exists
   Input 1 to overwrite it, 2 to choose another name ([1:2], <cr>=1):

   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):

   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1):

   Input 1 for default options, 2 otherwise ([1:2], <cr>=1): 2

   Input 1 for simple, 2 for nearest-neighbor, 3 for kernel method ([1:3], <cr>=1): 3

   Input 1 for univariate, 2 for bivariate preference ([1:2], <cr>=2):

   Input 1 for interaction tests, 2 to skip them ([1:2], <cr>=1):
4.3 Kernel discriminant models

Input 0 to specify tree with fixed no. of nodes, 1 to prune by CV, 2 by test sample, 3 for no pruning.

Input name of data description file (max 100 characters);
enclose with matching quotes if it has spaces: rhcdsc1.txt

rhcdsc1.txt
Reading data description file ...
Training sample file: rhcdata.txt
Missing value code: NA
Records in data file start on line 2
23 N variables changed to S
D variable is swang1
Reading data file ...
Number of records in data file: 5735
Length of longest entry in data file: 19
Checking for missing values ...
Finished checking
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Recoding D values to integers
Finished recoding
Number of classes: 2
Assigning integer codes to values of 30 categorical variables
Finished assigning codes to 10 categorical variables
Finished assigning codes to 20 categorical variables
Finished assigning codes to 30 categorical variables
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...
Finished processing 5000 of 5735 observations
Data checks complete
Creating missing value indicators ...
Rereading data ...

<table>
<thead>
<tr>
<th>Class</th>
<th>#Cases</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoRHC</td>
<td>3551</td>
<td>0.61918047</td>
</tr>
<tr>
<td>RHC</td>
<td>2184</td>
<td>0.38081953</td>
</tr>
</tbody>
</table>

Total #cases w/ #missing

<table>
<thead>
<tr>
<th>#cases miss. D ord. vals</th>
<th>#X-var</th>
<th>#N-var</th>
<th>#F-var</th>
<th>#S-var</th>
<th>#P-var</th>
<th>#M-var</th>
<th>#B-var</th>
<th>#C-var</th>
<th>#I-var</th>
</tr>
</thead>
<tbody>
<tr>
<td>5735</td>
<td>0</td>
<td>5157</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>30</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Number of cases used for training: 5735
Number of split variables: 53
Number of cases excluded due to 0 W or missing D variable: 0
Finished reading data file
4.3 Kernel discriminant models

Default number of cross-validations: 10
Input 1 to accept the default, 2 to change it ([1:2], <cr>=1):

Best tree may be chosen based on mean or median CV estimate
Input 1 for mean-based, 2 for median-based ([1:2], <cr>=1):

Input number of SEs for pruning ([0.00:1000.00], <cr>=0.25):
Choose 1 for estimated priors, 2 for equal priors, 3 to input priors from a file
Input 1, 2, or 3 ([1:3], <cr>=1):

Choose 1 for unit misclassification costs, 2 to input costs from a file
Input 1 or 2 ([1:2], <cr>=1):

Choose a split point selection method for numerical variables:
Choose 1 to use faster method based on sample quantiles
Choose 2 to use exhaustive search
Input 1 or 2 ([1:2], <cr>=2):

Default max. number of split levels: 15
Input 1 to accept this value, 2 to change it ([1:2], <cr>=1):

Default minimum node sample size is 57
Input 1 to use the default value, 2 to change it ([1:2], <cr>=1):

Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):

Input file name to store LaTeX code (use .tex as suffix): ker2.tex
ker2.tex
A file by that name exists
Input 1 to overwrite it, 2 to choose another name ([1:2], <cr>=1):

Input 1 to color terminal nodes, 2 otherwise ([1:2], <cr>=1):

Choose amount of detail in nodes of LaTeX tree diagram:
Input 0 for #errors, 1 for sample sizes, 2 for sample proportions, 3 for posterior probs, 4 for none
Input your choice ([0:4], <cr>=2):

You can store the variables and/or values used to split and fit in a file
Choose 1 to skip this step, 2 to store split and fit variables,
3 to store split variables and their values
Input your choice ([1:3], <cr>=1):

Input 2 to save fitted values and node IDs, 1 otherwise ([1:2], <cr>=2):

Input name of file to store node ID and fitted value of each case: ker2.fit

Wei-Yin Loh 49 GUIDE manual
ker2.fit
A file by that name exists
Input 1 to overwrite it, 2 to choose another name ([1:2], <cr>=1):

Input rank of top variable to split root node ([1:53], <cr>=1):

Input file is created!
Run GUIDE with the command: guide < ker2.in

4.3.2 Contents of ker2.out

Classification tree
Pruning by cross-validation
Data description file: rhcdsc1.txt
Training sample file: rhcdata.txt
Missing value code: NA
Records in data file start on line 2
23 N variables changed to S
D variable is swang1
Number of records in data file: 5735
Length of longest entry in data file: 19
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Number of classes: 2
Training sample class proportions of D variable swang1:

<table>
<thead>
<tr>
<th>Class</th>
<th>#Cases</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoRHC</td>
<td>3551</td>
<td>0.61918047</td>
</tr>
<tr>
<td>RHC</td>
<td>2184</td>
<td>0.38081953</td>
</tr>
</tbody>
</table>

Summary information for training sample of size 5735
d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators),
s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
m=missing-value flag variable, w=weight

<table>
<thead>
<tr>
<th>#Codes/</th>
<th>Levels/</th>
<th>Column</th>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>cat1</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>cat2</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>ca</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>cardiohx</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>58</td>
<td>ortho</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>59</td>
<td>addld3p</td>
<td>s</td>
<td>0.000</td>
<td>7.000</td>
<td>4296</td>
</tr>
</tbody>
</table>

Wei-Yin Loh 50 GUIDE manual
4.3 Kernel discriminant models

<table>
<thead>
<tr>
<th>urin1</th>
<th>race</th>
<th>income</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>c</td>
<td>c</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#cases w/ #missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>#cases</td>
</tr>
<tr>
<td>5735</td>
</tr>
<tr>
<td>#P-var</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

Number of cases used for training: 5735
Number of split variables: 53
Number of cases excluded due to 0 W or missing D variable: 0

Constant fitted to cases with missing values in regessor variables
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE's for pruned tree: 0.2500

Kernel density node models
Bivariate preference
Estimated priors
Unit misclassification costs
Bivariate split highest priority
Interaction splits 2nd priority; no linear splits
Split values for N and S variables based on exhaustive search
Maximum number of split levels: 15
Minimum node sample size: 57
Non-univariate split at root node

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Nodes</th>
<th>Mean Cost</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median Cost</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>76</td>
<td>3.170E-01</td>
<td>6.144E-03</td>
<td>7.391E-03</td>
<td>3.206E-01</td>
<td>1.024E-02</td>
</tr>
<tr>
<td>2</td>
<td>75</td>
<td>3.170E-01</td>
<td>6.144E-03</td>
<td>7.391E-03</td>
<td>3.206E-01</td>
<td>1.024E-02</td>
</tr>
<tr>
<td>3</td>
<td>74</td>
<td>3.170E-01</td>
<td>6.144E-03</td>
<td>7.391E-03</td>
<td>3.206E-01</td>
<td>1.024E-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>44+</td>
<td>15</td>
<td>3.065E-01</td>
<td>6.088E-03</td>
<td>5.357E-03</td>
<td>3.075E-01</td>
<td>7.515E-03</td>
</tr>
<tr>
<td>45+</td>
<td>14</td>
<td>3.039E-01</td>
<td>6.074E-03</td>
<td>4.918E-03</td>
<td>3.025E-01</td>
<td>5.966E-03</td>
</tr>
<tr>
<td>46++</td>
<td>9</td>
<td>3.043E-01</td>
<td>6.076E-03</td>
<td>5.104E-03</td>
<td>3.034E-01</td>
<td>4.222E-03</td>
</tr>
<tr>
<td>47**</td>
<td>7</td>
<td>3.039E-01</td>
<td>6.074E-03</td>
<td>5.098E-03</td>
<td>3.092E-01</td>
<td>7.207E-03</td>
</tr>
<tr>
<td>48</td>
<td>6</td>
<td>3.107E-01</td>
<td>6.111E-03</td>
<td>4.164E-03</td>
<td>3.121E-01</td>
<td>4.682E-03</td>
</tr>
<tr>
<td>49</td>
<td>5</td>
<td>3.180E-01</td>
<td>6.150E-03</td>
<td>5.979E-03</td>
<td>3.145E-01</td>
<td>8.560E-03</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>3.229E-01</td>
<td>6.175E-03</td>
<td>4.475E-03</td>
<td>3.194E-01</td>
<td>6.704E-03</td>
</tr>
<tr>
<td>51</td>
<td>3</td>
<td>3.236E-01</td>
<td>6.178E-03</td>
<td>4.577E-03</td>
<td>3.211E-01</td>
<td>7.707E-03</td>
</tr>
<tr>
<td>52</td>
<td>2</td>
<td>3.275E-01</td>
<td>6.197E-03</td>
<td>6.713E-03</td>
<td>3.211E-01</td>
<td>7.780E-03</td>
</tr>
<tr>
<td>53</td>
<td>1</td>
<td>3.688E-01</td>
<td>6.371E-03</td>
<td>2.637E-03</td>
<td>3.670E-01</td>
<td>2.864E-03</td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 7 terminal nodes
0-SE tree based on median is marked with + and has 14 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
** tree same as -- tree
* tree same as ** tree
* tree same as -- tree

Following tree is based on mean CV with naive SE estimate (**)

Structure of final tree. Each terminal node is marked with a T.

<table>
<thead>
<tr>
<th>Node label</th>
<th>Total cases</th>
<th>Train cases</th>
<th>Predicted class</th>
<th>Node cost</th>
<th>Split variable followed by (+)fit variable(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5735</td>
<td>5735</td>
<td>NoRHC</td>
<td>3.643E-01</td>
<td>cat1 +cat1 +pafi1</td>
</tr>
<tr>
<td>2</td>
<td>1683</td>
<td>1683</td>
<td>RHC</td>
<td>4.225E-01</td>
<td>adld3p +adld3p +pafi1</td>
</tr>
<tr>
<td>4</td>
<td>1183</td>
<td>1183</td>
<td>RHC</td>
<td>3.567E-01</td>
<td>wtkilo1 +wtkilo1 +pafi1</td>
</tr>
<tr>
<td>8T</td>
<td>452</td>
<td>452</td>
<td>NoRHC</td>
<td>3.540E-01</td>
<td>pafi1 +pafi1 +hema1</td>
</tr>
<tr>
<td>9T</td>
<td>731</td>
<td>731</td>
<td>RHC</td>
<td>3.010E-01</td>
<td>pafi1 +pafi1 +meanbp1</td>
</tr>
<tr>
<td>5</td>
<td>500</td>
<td>500</td>
<td>NoRHC</td>
<td>4.160E-01</td>
<td>card +card +meanbp1</td>
</tr>
<tr>
<td>10</td>
<td>345</td>
<td>345</td>
<td>NoRHC</td>
<td>3.420E-01</td>
<td>pot1 +pot1 +meanbp1</td>
</tr>
<tr>
<td>20T</td>
<td>181</td>
<td>181</td>
<td>RHC</td>
<td>2.928E-01</td>
<td>meanbp1 +meanbp1 +resp1</td>
</tr>
<tr>
<td>21T</td>
<td>164</td>
<td>164</td>
<td>NoRHC</td>
<td>2.683E-01</td>
<td>meanbp1 +meanbp1 +edu</td>
</tr>
<tr>
<td>11T</td>
<td>155</td>
<td>155</td>
<td>NoRHC</td>
<td>3.677E-01</td>
<td>resp1 +resp1</td>
</tr>
<tr>
<td>3</td>
<td>4052</td>
<td>4052</td>
<td>NoRHC</td>
<td>2.850E-01</td>
<td>pafi1 +pafi1 +crea1</td>
</tr>
<tr>
<td>6T</td>
<td>1281</td>
<td>1281</td>
<td>NoRHC</td>
<td>3.599E-01</td>
<td>aps1 +aps1 +resp1</td>
</tr>
<tr>
<td>7T</td>
<td>2771</td>
<td>2771</td>
<td>NoRHC</td>
<td>2.324E-01</td>
<td>meanbp1 +meanbp1 +crea1</td>
</tr>
</tbody>
</table>

Number of terminal nodes of final tree: 7
Total number of nodes of final tree: 13
Second best split variable (based on interaction test) at root node is pafi1

Classification tree:
For categorical variable splits, values not in training data go to the right

Node 1: cat1 = "CHF", "MOSF w/Sepsis"
Node 2: adld3p = NA
  Node 4: wtkilo1 <= 70.249970
  Node 8: Mean cost = 0.35398230
  Node 4: wtkilo1 > 70.249970 or NA
  Node 9: Mean cost = 0.30095759
Node 2: adld3p /= NA
  Node 5: card = "Yes"
  Node 10: pot1 <= 3.9499510
  Node 20: Mean cost = 0.29281768
4.3 Kernel discriminant models

Node 10: pot1 > 3.9499510 or NA
Node 21: Mean cost = 0.26829268
Node 5: card /= "Yes"
Node 11: Mean cost = 0.36774194
Node 1: cat1 /= "CHF", "MOSF w/Sepsis"
Node 3: pafi1 <= 141.85938
Node 6: Mean cost = 0.35987510
Node 3: pafi1 > 141.85938 or NA
Node 7: Mean cost = 0.23240707

**********************************************************************
Predictor means below are means of cases with no missing values.

Node 1: Intermediate node
A case goes into Node 2 if cat1 = "CHF", "MOSF w/Sepsis"
cat1 mode = ARF
pafi1 mean = 222.27371

<table>
<thead>
<tr>
<th>Class</th>
<th>Number</th>
<th>Posterior pafi1</th>
<th>Cat1 mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoRHC</td>
<td>3551</td>
<td>0.6192E+00</td>
<td>1.4868E-02</td>
</tr>
<tr>
<td>RHC</td>
<td>2184</td>
<td>0.3808E+00</td>
<td>1.2981E-02</td>
</tr>
</tbody>
</table>

Number of training cases misclassified = 2089
If node model is inapplicable due to missing values, predicted class is "NoRHC"

----------------------------
Node 2: Intermediate node
A case goes into Node 4 if adld3p = NA
adld3p mean = 1.234000
pafi1 mean = 249.20858

<table>
<thead>
<tr>
<th>Class</th>
<th>Number</th>
<th>Posterior pafi1</th>
<th>adld3p mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoRHC</td>
<td>774</td>
<td>0.4599E+00</td>
<td>7.6307E+01</td>
</tr>
<tr>
<td>RHC</td>
<td>909</td>
<td>0.5401E+00</td>
<td>6.8628E+01</td>
</tr>
</tbody>
</table>

Number of training cases misclassified = 711
If node model is inapplicable due to missing values, predicted class is "RHC"

----------------------------
Node 4: Intermediate node
A case goes into Node 8 if wtkilo1 <= 70.249970
wtkilo1 mean = 77.015038
pafi1 mean = 231.38524

<table>
<thead>
<tr>
<th>Class</th>
<th>Number</th>
<th>Posterior pafi1</th>
<th>wtkilo1 mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoRHC</td>
<td>488</td>
<td>0.4125E+00</td>
<td>9.4062E+01</td>
</tr>
<tr>
<td>RHC</td>
<td>695</td>
<td>0.5875E+00</td>
<td>7.1161E+01</td>
</tr>
</tbody>
</table>

Number of training cases misclassified = 422
If node model is inapplicable due to missing values, predicted class is "RHC"
4.3 Kernel discriminant models

Node 8: Terminal node
pafi1 mean = 244.88658
hema1 mean = 30.163116

Bandwidth
Class Number Posterior pafi1 hema1 Correlation
NoRHC 238 0.5265E+00 1.1248E+02 5.8918E+00 -0.1432
RHC 214 0.4735E+00 9.2951E+01 3.9603E+00 0.0123

Number of training cases misclassified = 160
If node model is inapplicable due to missing values, predicted class is "NoRHC"

------------------------------

Node 6: Terminal node
aps1 mean = 60.373927
resp1 mean = 30.854487

Bandwidth
Class Number Posterior aps1 resp1 Correlation
NoRHC 661 0.5160E+00 1.1125E+01 8.1589E+00 0.3789
RHC 620 0.4840E+00 1.2805E+01 9.8982E+00 0.3688

Number of training cases misclassified = 461
If node model is inapplicable due to missing values, predicted class is "NoRHC"

------------------------------

Node 7: Terminal node
meanbp1 mean = 85.416758
crea1 mean = 1.8756021

Bandwidth
Class Number Posterior meanbp1 crea1 Correlation
NoRHC 2116 0.7636E+00 2.0881E+01 4.0068E-01 -0.0610
RHC 655 0.2364E+00 2.3948E+01 8.6122E-01 -0.0970

Number of training cases misclassified = 644
If node model is inapplicable due to missing values, predicted class is "NoRHC"

------------------------------

Classification matrix for training sample:
Predicted class
NoRHC RHC
NoRHC 3003 1091
RHC 548 1093
Total 3551 2184

Number of cases used for tree construction: 5735
Number misclassified: 1639
Resubstitution estimate of mean misclassification cost: 0.28578901

Observed and fitted values are stored in ker2.fit
LaTeX code for tree is in ker2.tex
The kernel discriminant tree is shown in Figure 4. The row with two asterisks (**) in the output file ker2.out shows that the tree has 6 terminal nodes and a cross-validation estimate of misclassification cost of 0.3165. Unlike the default and linear-split trees, the class of each observation in a terminal node is predicted based on kernel discrimination and therefore is not constant within the node. The file ker2.fit contains the terminal node number, estimated posteriors class probabilities, and observed and predicted class of each observation. Following are the first 5 lines.

<table>
<thead>
<tr>
<th>train</th>
<th>node</th>
<th>&quot;P(NoRHC)&quot;</th>
<th>&quot;P(RHC)&quot;</th>
<th>observed</th>
<th>predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>6</td>
<td>0.47392</td>
<td>0.52608</td>
<td>&quot;NoRHC&quot;</td>
<td>&quot;RHC&quot;</td>
</tr>
<tr>
<td>y</td>
<td>8</td>
<td>0.45177</td>
<td>0.54823</td>
<td>&quot;RHC&quot;</td>
<td>&quot;RHC&quot;</td>
</tr>
<tr>
<td>y</td>
<td>7</td>
<td>0.60626</td>
<td>0.39374</td>
<td>&quot;RHC&quot;</td>
<td>&quot;NoRHC&quot;</td>
</tr>
<tr>
<td>y</td>
<td>7</td>
<td>0.77436</td>
<td>0.22564</td>
<td>&quot;NoRHC&quot;</td>
<td>&quot;NoRHC&quot;</td>
</tr>
<tr>
<td>y</td>
<td>9</td>
<td>0.32030</td>
<td>0.67970</td>
<td>&quot;RHC&quot;</td>
<td>&quot;RHC&quot;</td>
</tr>
</tbody>
</table>

Figure 5 shows plots of the data and the predicted values in terminal node 8 of the tree in the space of variables hema1 and pafi1 selected by GUIDE (see the information for these terminal nodes in ker2.out). The R code for making the plot is below.

```R
par(mfrow=c(1,2),pty="s",cex.lab=1.2,cex.axis=1.2,cex.main=1.5)
z1 <- read.table("ker2.fit",header=TRUE)
leg.txt <- c("NoRHC","RHC")
leg.col <- c("red","blue")
leg.pch <- rep(1,2)
gp <- z1$node == 8
x <- z0$pafi1[gp]
y <- z0$hema1[gp]
classv <- z0$swang1[gp]
plot(x,y,ylab="hema1",xlab="pafi1",type="n")
g1 <- classv == "NoRHC"
points(x[g1],y[g1],pch=leg.pch[1],col=leg.col[1])
points(x[!g1],y[!g1],pch=leg.pch[2],col=leg.col[2])
legend("topright",legend=leg.txt,col=leg.col,pch=leg.pch,cex=1.2)
title("Observed values in Node 8")
plot(x,y,ylab="hema1",xlab="pafi1",type="n")
pred <- z1$predicted[gp]
g1 <- pred == "NoRHC"
points(x[g1],y[g1],pch=leg.pch[1],col=leg.col[1])
points(x[!g1],y[!g1],pch=leg.pch[2],col=leg.col[2])
legend("topright",legend=leg.txt,col=leg.col,pch=leg.pch,cex=1.2)
title("Predicted values in Node 8")
```
4.3 Kernel discriminant models

4. CLASSIFICATION: RHC DATA

Figure 4: GUIDE v.41.0 0.250-SE classification tree for predicting \texttt{swang1} using bivariate kernel discriminant node models, estimated priors and unit misclassification costs. At each split, an observation goes to the left branch if and only if the condition is satisfied. $S_1 = \{\text{CHF, MOSF w/Sepsis}\}$. Predicted classes and sample sizes (in italics) printed below terminal nodes; class sample proportion for \texttt{swang1} = RHC beside nodes. Second best split variable (based on interaction test) at root node is \texttt{pafi1}.
4.4 Nearest-neighbor models

Yet another way to reduce the size of the default classification tree is to fit a nearest-neighbor model in each node. GUIDE can use univariate or bivariate nearest neighbors. We show this with bivariate neighbors here.

4.4.1 Input file generation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: nn2.in
   Input 1 for model fitting, 2 for importance or DIF scoring,
   3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: nn2.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1):
   Input 1 for default options, 2 otherwise ([1:2], <cr>=1): 2
   Input 1 for simple, 2 for nearest-neighbor, 3 for kernel method ([1:3], <cr>=1): 2
   Input 1 for univariate, 2 for bivariate preference ([1:2], <cr>=2):
   Input 1 for interaction tests, 2 to skip them ([1:2], <cr>=1):
   Input 0 to specify tree with fixed no. of nodes, 1 to prune by CV, 2 by test sample,
4.4 Nearest-neighbor models

3 for no pruning ([0:3], <cr>=1):
Input name of data description file (max 100 characters);
enclave with matching quotes if it has spaces: rhcdsc1.txt
Reading data description file ...
Training sample file: rhcdata.txt
Missing value code: NA
Records in data file start on line 2
23 N variables changed to S
D variable is swang1
Reading data file ...
Number of records in data file: 5735
Length of longest entry in data file: 19
Checking for missing values ...
Finished checking
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Recoding D values to integers
Finished recoding
Number of classes: 2
Assigning integer codes to values of 30 categorical variables
Finished assigning codes to 10 categorical variables
Finished assigning codes to 20 categorical variables
Finished assigning codes to 30 categorical variables
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...
Finished processing 5000 of 5735 observations
Data checks complete
Creating missing value indicators ...
Rereading data ...
Class  #Cases  Proportion
NoRHC 3551 0.61918047
RHC 2184 0.38081953
Total  #cases w/ #missing
#cases miss. D ord. vals #X-var #N-var #F-var #S-var
5735 0 5157 10 0 0 0 23
#P-var #M-var #B-var #C-var #I-var
0 0 0 30 0
Number of cases used for training: 5735
Number of split variables: 53
Number of cases excluded due to 0 W or missing D variable: 0
Finished reading data file
Default number of cross-validations: 10
Input 1 to accept the default, 2 to change it ([1:2], <cr>=1):
Best tree may be chosen based on mean or median CV estimate

Wei-Yin Loh  58  GUIDE manual
4.4 Nearest-neighbor models

Input 1 for mean-based, 2 for median-based ([1:2], <cr>=1):
Input number of SEs for pruning ([0.00:1000.00], <cr>=0.25):
Choose 1 for estimated priors, 2 for equal priors, 3 to input priors from a file
Input 1, 2, or 3 ([1:3], <cr>=1):
Choose 1 for unit misclassification costs, 2 to input costs from a file
Input 1 or 2 ([1:2], <cr>=1):
Choose a split point selection method for numerical variables:
Choose 1 to use faster method based on sample quantiles
Choose 2 to use exhaustive search
Input 1 or 2 ([1:2], <cr>=2):
Default max. number of split levels: 15
Input 1 to accept this value, 2 to change it ([1:2], <cr>=1):
Default minimum node sample size is 57
Input 1 to use the default value, 2 to change it ([1:2], <cr>=1):
Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):
Input file name to store LaTeX code (use .tex as suffix): nn2.tex
Input 1 to color terminal nodes, 2 otherwise ([1:2], <cr>=1):
Choose amount of detail in nodes of LaTeX tree diagram:
Input 0 for #errors, 1 for sample sizes, 2 for sample proportions, 3 for posterior probs, 4 for nothing
Input your choice ([0:4], <cr>=2):
you can store the variables and/or values used to split and fit in a file
choose 1 to skip this step, 2 to store split and fit variables,
3 to store split variables and their values
Input your choice ([1:3], <cr>=1):
input 2 to save fitted values and node ids, 1 otherwise ([1:2], <cr>=2):
Input rank of top variable to split root node ([1:53], <cr>=1):
Input file is created!
Run GUIDE with the command: guide < nn2.in

4.4.2 Contents of nn2.out

Classification tree
Pruning by cross-validation
Data description file: rhcdsacl.txt
Training sample file: rhcdata.txt
Missing value code: NA
Records in data file start on line 2
23 N variables changed to S
D variable is swang1
Number of records in data file: 5735
Length of longest entry in data file: 19
Missing values found among categorical variables
Separate categories will be created for missing categorical variables

Wei-Yin Loh 59 GUIDE manual
Missing values found among non-categorical variables
Number of classes: 2
Training sample class proportions of D variable swang1:

<table>
<thead>
<tr>
<th>Class</th>
<th>#Cases</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoRHC</td>
<td>3551</td>
<td>0.61918047</td>
</tr>
<tr>
<td>RHC</td>
<td>2184</td>
<td>0.38081953</td>
</tr>
</tbody>
</table>

Summary information for training sample of size 5735

<table>
<thead>
<tr>
<th>#Codes/Levels/Column Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 cat1 c</td>
<td></td>
<td></td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>3 cat2 c</td>
<td></td>
<td></td>
<td>6</td>
<td>4535</td>
</tr>
<tr>
<td>4 ca c</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>10 cardiohx c</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>11 chfhx c</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>12 dementhx c</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>13 psychhx c</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>14 chrpulhx c</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>15 renalhx c</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>16 liverhx c</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>17 gibledhx c</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>18 malighx c</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>19 immunhx c</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>20 transhx c</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>21 amihx c</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>22 age s</td>
<td>18.04</td>
<td>101.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23 sex c</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>24 edu s</td>
<td>0.000</td>
<td>30.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25 surv2md1 s</td>
<td>0.000</td>
<td>0.9620</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26 das2d3pc s</td>
<td>11.00</td>
<td>33.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29 aps1 s</td>
<td>3.000</td>
<td>147.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30 scoma1 s</td>
<td>0.000</td>
<td>100.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31 meanbp1 s</td>
<td>10.00</td>
<td>259.0</td>
<td></td>
<td>80</td>
</tr>
<tr>
<td>32 wblc1 s</td>
<td>0.000</td>
<td>192.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33 hrt1 s</td>
<td>8.000</td>
<td>250.0</td>
<td></td>
<td>159</td>
</tr>
<tr>
<td>34 resp1 s</td>
<td>2.000</td>
<td>100.0</td>
<td></td>
<td>136</td>
</tr>
<tr>
<td>35 temp1 s</td>
<td>27.00</td>
<td>43.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36 pafi1 s</td>
<td>11.60</td>
<td>937.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37 alb1 s</td>
<td>0.3000</td>
<td>29.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>38 hemal s</td>
<td>2.000</td>
<td>66.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39 bilii1 s</td>
<td>0.9999E-01</td>
<td>58.20</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.4 Nearest-neighbor models

40 crea1 s 0.9999E-01 25.10
41 sod1 s 101.0 178.0
42 pot1 s 1.100 11.90
43 paco21 s 1.000 156.0
44 ph1 s 6.579 7.770
45 swang1 d 2
46 wtkilo1 s 19.50 244.0 515
47 dnr1 c 2
48 ninsclas c 6
49 resp c 2
50 card c 2
51 neuro c 2
52 gastr c 2
53 renal c 2
54 meta c 2
55 hema c 2
56 seps c 2
57 trauma c 2
58 ortho c 2
59 addl3p s 0.000 7.000 4296
60 urin1 s 0.000 9000. 3028
61 race c 3
62 income c 4

Total #cases w/ #missing
#cases miss. D ord. vals #X-var #N-var #F-var #S-var
5735 0 5157 10 0 0 23
#P-var #M-var #B-var #C-var #I-var
0 0 0 0 30 0

Number of cases used for training: 5735
Number of split variables: 53
Number of cases excluded due to 0 W or missing D variable: 0

Constant fitted to cases with missing values in regressor variables
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE’s for pruned tree: 0.2500

Nearest-neighbor node models
Bivariate preference
Estimated priors
Unit misclassification costs
Bivariate split highest priority
Interaction splits 2nd priority; no linear splits
Split values for N and S variables based on exhaustive search
Maximum number of split levels: 15
Minimum node sample size: 57
Non-univariate split at root node

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Nodes</th>
<th>Mean Cost</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median Cost</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>76</td>
<td>3.151E-01</td>
<td>6.134E-03</td>
<td>6.384E-03</td>
<td>3.188E-01</td>
<td>1.012E-02</td>
</tr>
<tr>
<td>2</td>
<td>75</td>
<td>3.151E-01</td>
<td>6.134E-03</td>
<td>6.384E-03</td>
<td>3.188E-01</td>
<td>1.012E-02</td>
</tr>
<tr>
<td>3</td>
<td>74</td>
<td>3.151E-01</td>
<td>6.134E-03</td>
<td>6.384E-03</td>
<td>3.188E-01</td>
<td>1.012E-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>39</td>
<td>3.151E-01</td>
<td>6.134E-03</td>
<td>6.384E-03</td>
<td>3.188E-01</td>
<td>1.012E-02</td>
</tr>
<tr>
<td>37*</td>
<td>38</td>
<td>3.151E-01</td>
<td>6.134E-03</td>
<td>6.384E-03</td>
<td>3.188E-01</td>
<td>1.012E-02</td>
</tr>
<tr>
<td>38</td>
<td>36</td>
<td>3.153E-01</td>
<td>6.135E-03</td>
<td>5.373E-03</td>
<td>3.139E-01</td>
<td>8.226E-03</td>
</tr>
<tr>
<td>39++</td>
<td>35</td>
<td>3.153E-01</td>
<td>6.135E-03</td>
<td>5.373E-03</td>
<td>3.139E-01</td>
<td>8.226E-03</td>
</tr>
<tr>
<td>40</td>
<td>34</td>
<td>3.163E-01</td>
<td>6.141E-03</td>
<td>6.259E-03</td>
<td>3.173E-01</td>
<td>9.416E-03</td>
</tr>
<tr>
<td>42</td>
<td>31</td>
<td>3.163E-01</td>
<td>6.141E-03</td>
<td>6.111E-03</td>
<td>3.173E-01</td>
<td>8.898E-03</td>
</tr>
<tr>
<td>43++</td>
<td>29</td>
<td>3.163E-01</td>
<td>6.141E-03</td>
<td>6.111E-03</td>
<td>3.173E-01</td>
<td>8.898E-03</td>
</tr>
<tr>
<td>46</td>
<td>17</td>
<td>3.194E-01</td>
<td>6.157E-03</td>
<td>5.933E-03</td>
<td>3.191E-01</td>
<td>9.860E-03</td>
</tr>
<tr>
<td>48</td>
<td>15</td>
<td>3.182E-01</td>
<td>6.151E-03</td>
<td>5.834E-03</td>
<td>3.240E-01</td>
<td>8.063E-03</td>
</tr>
<tr>
<td>49</td>
<td>14</td>
<td>3.184E-01</td>
<td>6.152E-03</td>
<td>5.997E-03</td>
<td>3.261E-01</td>
<td>8.891E-03</td>
</tr>
<tr>
<td>50</td>
<td>9</td>
<td>3.184E-01</td>
<td>6.152E-03</td>
<td>5.997E-03</td>
<td>3.261E-01</td>
<td>8.891E-03</td>
</tr>
<tr>
<td>51</td>
<td>7</td>
<td>3.217E-01</td>
<td>6.168E-03</td>
<td>5.279E-03</td>
<td>3.226E-01</td>
<td>6.723E-03</td>
</tr>
<tr>
<td>52</td>
<td>5</td>
<td>3.250E-01</td>
<td>6.185E-03</td>
<td>6.166E-03</td>
<td>3.243E-01</td>
<td>1.047E-02</td>
</tr>
<tr>
<td>53</td>
<td>1</td>
<td>3.439E-01</td>
<td>6.272E-03</td>
<td>4.168E-03</td>
<td>3.458E-01</td>
<td>7.691E-03</td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 38 terminal nodes
0-SE tree based on median is marked with + and has 35 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
** tree same as -- tree
+ tree same as ++ tree

Following tree is based on mean CV with naive SE estimate (**)

Structure of final tree. Each terminal node is marked with a T.

Node cost is node misclassification cost divided by number of training cases

<table>
<thead>
<tr>
<th>Node label</th>
<th>Total cases</th>
<th>Train cases</th>
<th>Predicted cases</th>
<th>Split variable followed by (+)fit variable(s)</th>
<th>Node cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5735</td>
<td>5735</td>
<td>NoRHC</td>
<td>2.961E-01 cat1 +cat1 +pafi1</td>
<td>2.961E-01</td>
</tr>
<tr>
<td>2</td>
<td>1683</td>
<td>1683</td>
<td>RHC</td>
<td>4.029E-01 adld3p +adld3p +pafi1</td>
<td>4.029E-01</td>
</tr>
<tr>
<td>4</td>
<td>1183</td>
<td>1183</td>
<td>RHC</td>
<td>3.271E-01 wtkilo1 +wtkilo1 +pafi1</td>
<td>3.271E-01</td>
</tr>
<tr>
<td>8</td>
<td>452</td>
<td>452</td>
<td>NoRHC</td>
<td>2.942E-01 pafi1 +pafi1 +hema1</td>
<td>2.942E-01</td>
</tr>
</tbody>
</table>
### 4.4 Nearest-neighbor models

<table>
<thead>
<tr>
<th>Time (h)</th>
<th>Nearest-neighbor</th>
<th>Value</th>
<th>Feature 1</th>
<th>Feature 2</th>
<th>Feature 3</th>
<th>Class</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>16T</td>
<td>257</td>
<td>RHC</td>
<td>hema1</td>
<td>hema1</td>
<td>ph1</td>
<td>2.646E-01</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>195</td>
<td>NoRHC</td>
<td>age</td>
<td>age</td>
<td></td>
<td>2.872E-01</td>
<td></td>
</tr>
<tr>
<td>34T</td>
<td>137</td>
<td>NoRHC</td>
<td>das2d3pc</td>
<td>das2d3pc</td>
<td></td>
<td>3.139E-01</td>
<td></td>
</tr>
<tr>
<td>35T</td>
<td>58</td>
<td>NoRHC</td>
<td></td>
<td></td>
<td></td>
<td>1.034E-01</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>731</td>
<td>RHC</td>
<td>pafii</td>
<td>pafii</td>
<td>meanbp1</td>
<td>2.791E-01</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>420</td>
<td>RHC</td>
<td>sex</td>
<td>sex</td>
<td></td>
<td>2.619E-01</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>300</td>
<td>RHC</td>
<td>resp1</td>
<td>resp1</td>
<td>edu</td>
<td>2.233E-01</td>
<td></td>
</tr>
<tr>
<td>72T</td>
<td>90</td>
<td>RHC</td>
<td></td>
<td></td>
<td></td>
<td>6.667E-02</td>
<td></td>
</tr>
<tr>
<td>73</td>
<td>210</td>
<td>RHC</td>
<td>edu</td>
<td>edu</td>
<td></td>
<td>2.905E-01</td>
<td></td>
</tr>
<tr>
<td>146T</td>
<td>138</td>
<td>RHC</td>
<td>aps1</td>
<td>aps1</td>
<td></td>
<td>2.826E-01</td>
<td></td>
</tr>
<tr>
<td>147T</td>
<td>72</td>
<td>RHC</td>
<td></td>
<td></td>
<td></td>
<td>1.806E-01</td>
<td></td>
</tr>
<tr>
<td>37T</td>
<td>120</td>
<td>RHC</td>
<td>crea1</td>
<td>crea1</td>
<td></td>
<td>3.000E-01</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>311</td>
<td>RHC</td>
<td>meanbp1</td>
<td>meanbp1</td>
<td></td>
<td>2.990E-01</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>237</td>
<td>RHC</td>
<td>resp1</td>
<td>resp1</td>
<td></td>
<td>3.418E-01</td>
<td></td>
</tr>
<tr>
<td>76T</td>
<td>92</td>
<td>RHC</td>
<td></td>
<td></td>
<td></td>
<td>2.609E-01</td>
<td></td>
</tr>
<tr>
<td>77T</td>
<td>145</td>
<td>RHC</td>
<td>age</td>
<td>age</td>
<td></td>
<td>3.103E-01</td>
<td></td>
</tr>
<tr>
<td>39T</td>
<td>74</td>
<td>NoRHC</td>
<td>card</td>
<td>card</td>
<td>meanbp1</td>
<td>2.432E-01</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>500</td>
<td>NoRHC</td>
<td>pot1</td>
<td>pot1</td>
<td>meanbp1</td>
<td>3.280E-01</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>345</td>
<td>NoRHC</td>
<td></td>
<td></td>
<td>meanbp1</td>
<td>3.072E-01</td>
<td></td>
</tr>
<tr>
<td>20T</td>
<td>181</td>
<td>RHC</td>
<td>meanbp1</td>
<td>meanbp1</td>
<td>resp1</td>
<td>2.652E-01</td>
<td></td>
</tr>
<tr>
<td>21T</td>
<td>164</td>
<td>RHC</td>
<td>meanbp1</td>
<td>meanbp1</td>
<td>edu</td>
<td>2.805E-01</td>
<td></td>
</tr>
<tr>
<td>11T</td>
<td>155</td>
<td>NoRHC</td>
<td>resp1</td>
<td>resp1</td>
<td></td>
<td>3.226E-01</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>4052</td>
<td>NoRHC</td>
<td>pafii</td>
<td>pafii</td>
<td>crea1</td>
<td>2.848E-01</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1281</td>
<td>NoRHC</td>
<td>aps1</td>
<td>aps1</td>
<td>resp1</td>
<td>3.052E-01</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>855</td>
<td>NoRHC</td>
<td></td>
<td></td>
<td></td>
<td>4.234E-01</td>
<td></td>
</tr>
<tr>
<td>24T</td>
<td>272</td>
<td>RHC</td>
<td>alb1</td>
<td>alb1</td>
<td>meanbp1</td>
<td>3.125E-01</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>583</td>
<td>NoRHC</td>
<td>resp</td>
<td>resp</td>
<td></td>
<td>3.588E-01</td>
<td></td>
</tr>
<tr>
<td>50T</td>
<td>182</td>
<td>NoRHC</td>
<td>edu</td>
<td>edu</td>
<td></td>
<td>3.462E-01</td>
<td></td>
</tr>
<tr>
<td>51T</td>
<td>401</td>
<td>NoRHC</td>
<td>immunhx</td>
<td>immunhx</td>
<td>temp1</td>
<td>2.693E-01</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>426</td>
<td>RHC</td>
<td>resp</td>
<td>resp</td>
<td>resp1</td>
<td>3.427E-01</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>224</td>
<td>RHC</td>
<td>resp1</td>
<td>resp1</td>
<td>age</td>
<td>3.080E-01</td>
<td></td>
</tr>
<tr>
<td>52T</td>
<td>139</td>
<td>RHC</td>
<td>ph1</td>
<td>ph1</td>
<td></td>
<td>2.302E-01</td>
<td></td>
</tr>
<tr>
<td>53T</td>
<td>85</td>
<td>NoRHC</td>
<td></td>
<td></td>
<td></td>
<td>3.059E-01</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>202</td>
<td>RHC</td>
<td>pac21</td>
<td>pac21</td>
<td></td>
<td>2.723E-01</td>
<td></td>
</tr>
<tr>
<td>54T</td>
<td>69</td>
<td>RHC</td>
<td></td>
<td></td>
<td></td>
<td>1.304E-01</td>
<td></td>
</tr>
<tr>
<td>55T</td>
<td>133</td>
<td>RHC</td>
<td>surv2md1</td>
<td>surv2md1</td>
<td></td>
<td>2.857E-01</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2771</td>
<td>NoRHC</td>
<td>meanbp1</td>
<td>meanbp1</td>
<td>crea1</td>
<td>2.317E-01</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1456</td>
<td>NoRHC</td>
<td>adld3p</td>
<td>adld3p</td>
<td>crea1</td>
<td>3.043E-01</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>1095</td>
<td>NoRHC</td>
<td>wtkilo1</td>
<td>wtkilo1</td>
<td>aps1</td>
<td>2.749E-01</td>
<td></td>
</tr>
<tr>
<td>56T</td>
<td>316</td>
<td>NoRHC</td>
<td>card</td>
<td>card</td>
<td>hema1</td>
<td>1.677E-01</td>
<td></td>
</tr>
<tr>
<td>57</td>
<td>779</td>
<td>NoRHC</td>
<td>dementhx</td>
<td>dementhx</td>
<td>crea1</td>
<td>3.389E-01</td>
<td></td>
</tr>
<tr>
<td>114</td>
<td>695</td>
<td>NoRHC</td>
<td>dnr1</td>
<td>dnr1</td>
<td>crea1</td>
<td>3.367E-01</td>
<td></td>
</tr>
<tr>
<td>228</td>
<td>617</td>
<td>NoRHC</td>
<td>pafii</td>
<td>pafii</td>
<td>crea1</td>
<td>2.966E-01</td>
<td></td>
</tr>
<tr>
<td>456T</td>
<td>262</td>
<td>RHC</td>
<td>cat2</td>
<td>cat2</td>
<td>crea1</td>
<td>2.595E-01</td>
<td></td>
</tr>
<tr>
<td>457</td>
<td>355</td>
<td>NoRHC</td>
<td>pac21</td>
<td>pac21</td>
<td>crea1</td>
<td>3.014E-01</td>
<td></td>
</tr>
<tr>
<td>914</td>
<td>190</td>
<td>NoRHC</td>
<td>ph1</td>
<td>ph1</td>
<td>crea1</td>
<td>2.684E-01</td>
<td></td>
</tr>
</tbody>
</table>
4.4 Nearest-neighbor models

Classification tree:
For categorical variable splits, values not in training data go to the right

Node 1: \text{cat1} = "CHF", "MOSF w/Sepsis"
Node 2: \text{adld3p} = NA
Node 4: \text{wtkilo1} \leq 70.249970
  Node 8: \text{pafi1} \leq 254.50000
    Node 16: Mean cost = 0.26459144
    Node 8: \text{pafi1} > 254.50000 or NA
    Node 17: \text{age} \leq 75.961460
      Node 34: Mean cost = 0.31386861
    Node 17: \text{age} > 75.961460 or NA
      Node 35: Mean cost = 0.10344828
Node 4: \text{wtkilo1} > 70.249970 or NA
Node 9: \text{pafi1} \leq 227.75000
  Node 18: \text{sex} = "Male"
    Node 36: \text{resp1} \leq 17.000000 or NA
      Node 72: Mean cost = 0.66666667E-1
    Node 36: \text{resp1} > 17.000000
      Node 73: \text{edu} \leq 12.410785
        Node 146: Mean cost = 0.28260870
      Node 73: \text{edu} > 12.410785 or NA
        Node 147: Mean cost = 0.18055556
    Node 18: \text{sex} /= "Male"
      Node 37: Mean cost = 0.30000000
Node 9: \text{pafi1} > 227.75000 or NA
Node 19: \text{meanbp1} \leq 106.50000 or NA
  Node 38: \text{resp1} \leq 25.500000 or NA
    Node 76: Mean cost = 0.26086957
  Node 38: \text{resp1} > 25.500000
    Node 77: Mean cost = 0.31034483
Node 19: \text{meanbp1} > 106.50000
  Node 39: Mean cost = 0.24324324

Second best split variable (based on interaction test) at root node is \text{pafi1}

Number of terminal nodes of final tree: 29
Total number of nodes of final tree: 57

Warning: tree very large, omitting node numbers in \LaTeX\ file
4.4 Nearest-neighbor models

Node 2: adld3p = NA
    Node 5: card = "Yes"
        Node 10: pot1 <= 3.9499510
            Node 20: Mean cost = 0.26519337
        Node 10: pot1 > 3.9499510 or NA
            Node 21: Mean cost = 0.28048780
        Node 5: card /= "Yes"
            Node 11: Mean cost = 0.32258065
    Node 1: cat1 /= "CHF", "MOSF w/Sepsis"
    Node 3: pafi1 <= 141.85938
        Node 6: aps1 <= 66.500000
            Node 12: card = "Yes"
                Node 24: Mean cost = 0.31250000
            Node 12: card /= "Yes"
                Node 25: resp = "No"
                    Node 50: Mean cost = 0.34615385
                Node 25: resp /= "No"
                    Node 51: Mean cost = 0.26932668
        Node 6: aps1 > 66.500000 or NA
        Node 13: resp = "Yes"
            Node 26: resp1 <= 41.000000
                Node 52: Mean cost = 0.23021583
            Node 26: resp1 > 41.000000 or NA
                Node 53: Mean cost = 0.30588235
        Node 13: resp /= "Yes"
            Node 27: paco21 <= 31.500000
                Node 54: Mean cost = 0.13043478
            Node 27: paco21 > 31.500000 or NA
                Node 55: Mean cost = 0.28571429
    Node 3: pafi1 > 141.85938 or NA
    Node 7: meanbp1 <= 69.500000 or NA
    Node 14: adld3p = NA
        Node 28: wtkilo1 <= 57.399995 or NA
            Node 56: Mean cost = 0.16772152
        Node 28: wtkilo1 > 57.399995
            Node 57: dementhx = "0"
                Node 114: dnr1 = "No"
                    Node 228: pafi1 <= 216.15625
                        Node 456: Mean cost = 0.25954198
                    Node 228: pafi1 > 216.15625 or NA
                        Node 457: paco21 <= 36.500000
                            Node 914: ph1 <= 7.4648440
                                Node 1828: Mean cost = 0.21600000
                            Node 914: ph1 > 7.4648440 or NA
                                Node 1829: Mean cost = 0.26153846
                            Node 457: paco21 > 36.500000 or NA
Nearest-neighbor models

Node 915: Mean cost = 0.26666667
Node 114: dnr1 /= "No"
  Node 229: Mean cost = 0.26923077
  Node 57: dementhx /= "0"
  Node 115: Mean cost = 0.21428571
  Node 14: adld3p /= NA
  Node 29: Mean cost = 0.18559557
  Node 7: meanbp1 > 69.500000
  Node 15: Mean cost = 0.16121673

***************************************************************
Predictor means below are means of cases with no missing values.

Node 1: Intermediate node
  A case goes into Node 2 if cat1 = "CHF", "MOD w/Sepsis"
  Number of nearest neighbors = 9
  cat1 mode = ARF
  pafi1 mean = 222.27371
  Class       Number  Posterior
  NoRHC       3551    0.6192E+00
  RHC         2184    0.3808E+00
  Number of training cases misclassified = 1698
  If node model is inapplicable due to missing values, predicted class is "NoRHC"

Node 2: Intermediate node
  A case goes into Node 4 if adld3p = NA
  Number of nearest neighbors = 8
  adld3p mean = 1.2340000 SD = 1.8633799
  pafi1 mean = 249.20858 SD = 104.96492
  correlation = 0.63530716E-1
  Class       Number  Posterior
  NoRHC       774     0.4599E+00
  RHC         909     0.5401E+00
  Number of training cases misclassified = 678
  If node model is inapplicable due to missing values, predicted class is "RHC"

N

Node 29: Terminal node
  Number of nearest neighbors = 6
  age mean = 62.145410
  card mode = No
  Class       Number  Posterior
  NoRHC       294     0.8144E+00
  RHC         67      0.1856E+00
  Number of training cases misclassified = 67

Wei-Yin Loh 66 GUIDE manual
4.4 Nearest-neighbor models

If node model is inapplicable due to missing values, predicted class is "NoRHC"

Node 15: Terminal node
Number of nearest neighbors = 8
hema1 mean = 33.662565
card mode = No
Class Number Posterior
NoRHC 1103 0.8388E+00
RHC 212 0.1612E+00
Number of training cases misclassified = 212
If node model is inapplicable due to missing values, predicted class is "NoRHC"

Classification matrix for training sample:

<table>
<thead>
<tr>
<th>Predicted</th>
<th>True class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NoRHC</td>
</tr>
<tr>
<td>NoRHC</td>
<td>3109</td>
</tr>
<tr>
<td>RHC</td>
<td>442</td>
</tr>
<tr>
<td>Total</td>
<td>3551</td>
</tr>
</tbody>
</table>

Number of cases used for tree construction: 5735
Number misclassified: 1330
Resubstitution estimate of mean misclassification cost: 0.23190933

Observed and fitted values are stored in nn2.fit
LaTeX code for tree is in nn2.tex

The nearest-neighbor density tree is shown in Figure 6. It is a supertree of the kernel discriminant tree in Figure 4. The row with two asterisks (**) in the output file nn2.out shows that the tree has 29 terminal nodes and a cross-validation estimate of misclassification cost of 0.3163. Unlike the default and linear-split trees, the class of each observation in a terminal node is predicted based on the classes of its neighbors and therefore is not constant within the node. Figure 7 shows plots of the data and the predicted values in terminal node 16 (leftmost node) of the tree in the space of variables hema1 and ph1 selected by GUIDE (see the information for these terminal nodes in nn2.out).
Table 1: Predicted classes and sample sizes (in italics) printed below terminal nodes; class sample proportion for swang1 = RHC beside nodes. Second best split variable (based on interaction test) at root node is pafi1.

Figure 6: GUIDE v.4.1.0.0-25-SE classification tree for predicting swang1 using bivariate nearest-neighbor node models; estimated priors and unit misclassification costs. At each split, an observation goes to the left branch if and only if the condition is satisfied. Symbol ‘≤’ stands for ‘≤ or missing’. S1 = {CHF, MOSF w/Sepsis}. Predicted classes and sample sizes (in italics) printed below terminal nodes; class sample proportion for swang1 = RHC beside nodes. Second best split variable (based on interaction test) at root node is pafi1.
Figure 7: Plots of observed and predicted values for data in node 16 of tree in Figure 6

File *nn2.fit* gives the terminal node number and observed and predicted classes of each observation in the data file. Below are the first 5 rows. The first column is "y" (for yes) or "n" (for no) if the observation is used or not used to train the model. Unlike the kernel discriminant model, there are no estimated posterior class probabilities.

<table>
<thead>
<tr>
<th>train</th>
<th>node</th>
<th>observed</th>
<th>predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>24</td>
<td>&quot;NoRHC&quot;</td>
<td>&quot;RHC&quot;</td>
</tr>
<tr>
<td>y</td>
<td>16</td>
<td>&quot;RHC&quot;</td>
<td>&quot;RHC&quot;</td>
</tr>
<tr>
<td>y</td>
<td>56</td>
<td>&quot;RHC&quot;</td>
<td>&quot;RHC&quot;</td>
</tr>
<tr>
<td>y</td>
<td>56</td>
<td>&quot;NoRHC&quot;</td>
<td>&quot;NoRHC&quot;</td>
</tr>
<tr>
<td>y</td>
<td>77</td>
<td>&quot;RHC&quot;</td>
<td>&quot;RHC&quot;</td>
</tr>
</tbody>
</table>
5 Missing-value flag variables: CE data

The Consumer Expenditure (CE) Survey is carried out by the Census Bureau for the Bureau of Labor Statistics (BLS). Conducted quarterly, the survey is a rotating panel survey that collects data on expenditures, income, and demographic characteristics of a sample of about 6000 consumer units (CUs) in the United States. After a CU is in the survey for four quarters, it is dropped and a new unit selected to replace it. The BLS defines CU and reference person of the CU as follows.

1. A CU consists of any of the following:
   (a) All members of a particular household who are related by blood, marriage, adoption, or other legal arrangements.
   (b) A person living alone or sharing a household with others or living as a roomer in a private home or lodging house or in permanent living quarters in a hotel or motel, but who is financially independent.
   (c) Two or more persons living together who use their incomes to make joint expenditure decisions. Financial independence is determined by spending behavior with regard to the three major expense categories: housing, food, and other living expenses. To be considered financially independent, the respondent must provide at least two of the three major expenditure categories, either entirely or in part.

2. A reference person of the CU is the first member mentioned by the respondent when asked “What are the names of all the persons living or staying here? Start with the name of the person or one of the persons who owns or rents the home.” It is with respect to this person that the relationship of the other CU members is determined.

The data in the file ce2021.txt consist of 3965 observations on 550 variables. They are extracted from the second, third and fourth quarters of 2021 and the first quarter of 2022 of the Interview part of the CE survey. For the purpose of illustration and because it is not possible to link CUs between quarters, each CU in the sample is treated as unique. Table 11 in the Appendix gives the names and definitions of some of the 550 variables and their missing-value rates.

About 20% of the variables are missing-value flags that give the reasons for missing values. Table 8 lists the flag codes. A variable takes value NA (nonresponse) if its flag variable code is A, B, or C. The names of flag variables are typically the same as their parents, except for the addition or substitution of an underscore. For
Table 8: Codes for missing-value flag variables

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Valid nonresponse; a response is not anticipated</td>
</tr>
<tr>
<td>B</td>
<td>Invalid nonresponse; nonresponse inconsistent with other data reported by CU</td>
</tr>
<tr>
<td>C</td>
<td>“Don’t know”, refusal, or other type of nonresponse</td>
</tr>
<tr>
<td>D</td>
<td>Valid unadjusted data value</td>
</tr>
<tr>
<td>T</td>
<td>Valid value topcoded or suppressed</td>
</tr>
</tbody>
</table>

Example, INTRDVX_ is the flag variable for INTRDVX (amount of income received from interest and dividends). In this dataset, INTRDVX_ has no B codes and records with A codes are removed. Thirty-seven percent of the records (1478) have INTRDVX_ = C.

A T flag code indicates that the value of a variable is “top-coded.” Top-coding is a method used by the BLS to protect the privacy of the respondents in the top 3 percent of the data. Usually, the reported values of the CUs in this group are replaced by their group mean. For example, below are the values of AGE2 (age of spouse) and AGE2_ in rows 112–117 of the data:

<table>
<thead>
<tr>
<th>AGE2</th>
<th>AGE2_</th>
</tr>
</thead>
<tbody>
<tr>
<td>112</td>
<td>29</td>
</tr>
<tr>
<td>113</td>
<td>87</td>
</tr>
<tr>
<td>114</td>
<td>NA</td>
</tr>
<tr>
<td>115</td>
<td>57</td>
</tr>
<tr>
<td>116</td>
<td>87</td>
</tr>
<tr>
<td>117</td>
<td>NA</td>
</tr>
</tbody>
</table>

Respondents 113 and 116 are topcoded and have their values equal to 87, the mean of the top 3 percent of AGE2. See https://www.bls.gov/cex/pumd_doc.htm for names of all the variables and Loh et al. (2019b, 2020) for an analysis of an earlier dataset.

Variable FINLWT21 is a sampling weight. For classification, GUIDE treats all observations with positive sampling weight equally; observations with non-positive weights are ignored in tree construction.

Missing-value flag variables are indicated by the letters “m” or “M” in the description file. To indicate to GUIDE to which variable is associated with which M variable, the latter must follow immediately after a B, C, N, P, or S variable in the description file. For example, the following lines from the file ce2021class.dsc indicate that DIRACC_ is the flag variable for DIRACC, AGE_REF_ is the flag variable for AGE_REF, and INCN_NW1 is the flag variable for INCNONW1.
ce2021.txt
NA
2
1 DIRACC n
2 DIRACC m
3 AGE_REF n
4 AGE_REF m
5 AGE2 n
6 AGE2 m
7 AS_COMP1 n
8 AS_COMP2 n
9 AS_COMP3 n
10 AS_COMP4 n
11 AS_COMP5 n
12 BATHRMQ n
13 BATHRMQ m
14 BEDROOMQ n
15 BEDROOMQ m
16 BLS_URBN n
17 BUILDING c
18 EARNCOMP c
19 EDUC_REF n
20 EDUC_REF m
21 EDUCA2 n
22 EDUCA2 m
50 INC_NONW1 c
51 INC_NONW1 m

A split on an N, P, or S variable that has an associated missing-value flag variable can take several forms. For example, a split on RETSURVX (retirement, survivor, or disability pensions in past 12 months) with flag variable RETS_RVX (which takes values A, C, D, and T) can take 7 forms:

1. RETS_RVX = A (only A flag values go left)
2. RETS_RVX = C (only C flag values go left)
3. RETSURVX = NA (all missing values go left)
5.1 Classification tree

4. RETSURVX ≤ c

5. RETSURVX ≤∗ c (the symbol “≤∗” means “≤ or is missing”)

6. RETSURVX ≤ c or RETS_RVX = A

7. RETSURVX ≤ c or RETS_RVX = C

Similarly, a split on a C variable such as INCNONW2 that has missing-value flag variable INCN_NW2 can take these forms (see Figure 14):

1. INCNONW2 in S (where S is a subset of values of INCNONW2)

2. INCNONW2 = NA

3. INCNONW2 in S or INCN_NW2 in S* (where S* is a subset of flag codes)

5.1 Classification tree

This section shows how to construct a classification tree for predicting INTRDVX_ using the description file ce2021class.dsc.

5.1.1 Input file generation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: class.in
   Input 1 for model fitting, 2 for importance or DIF scoring,
   3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: class.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1):
   Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
   Input name of data description file (max 100 characters); enclose with matching quotes if it has spaces: ce2021class.dsc
   Reading data description file ...
   Training sample file: ce2021.txt
   Missing value code: NA
   Records in data file start on line 2
   Number of M variables associated with C variables: 19
   384 N variables changed to S
   D variable is INTRDVX_
   Reading data file ...
Number of records in data file: 3965
Length of longest entry in data file: 11
Checking for missing values ...
Finished checking
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Recoding D values to integers
Finished recoding
Number of classes: 3
Finding number of levels of M variables associated with C variables ...
Assigning integer codes to values of 47 categorical variables
Finished assigning codes to 10 categorical variables
Finished assigning codes to 20 categorical variables
Finished assigning codes to 30 categorical variables
Finished assigning codes to 40 categorical variables
Associating missing values of N, P and S variables with M variable codes ...
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...
Data checks complete
Creating missing value indicators ...
Rereading data ...
Warning: S variable DIRACC is constant
Warning: S variable TOTHVHRP is constant
Warning: S variable TOTHVHRC is constant
Warning: S variable ROTHRFLC is constant
Warning: S variable WELFREBX is constant
Smallest positive weight: 1.0725E+03
Largest positive weight: 9.3902E+04
Class #Cases Proportion
C 1478 0.37276166
D 2431 0.61311475
T 56 0.01412358

Total #cases w/ #missing
#cases miss. D ord. vals #X-var #N-var #F-var #S-var
3965 0 3965 1 0 0 384

#P-var #M-var #B-var #C-var #I-var
0 116 0 47 0

Number of cases used for training: 3965
Number of split variables: 431
Number of cases excluded due to 0 W or missing D variable: 0
Finished reading data file
Warning: No interaction tests; too many predictor variables
Choose 1 for estimated priors, 2 for equal priors, 3 to input priors from a file
Input 1, 2, or 3 ([1:3], <cr>=1):
Choose 1 for unit misclassification costs, 2 to input costs from a file
Input 1 or 2 ([1:2], <cr>=1):
Warning: All positive weights treated as 1
Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):
Input file name to store LaTeX code (use .tex as suffix): class.tex
You can store the variables and/or values used to split and fit in a file
Choose 1 to skip this step, 2 to store split and fit variables,
3 to store split variables and their values
Input your choice ([1:3], <cr>=1):
Input 2 to save fitted values and node IDs, 1 otherwise ([1:2], <cr>=2):
Input name of file to store node ID and fitted value of each case: class.fit
Input 2 to write R function for predicting new cases, 1 otherwise ([1:2], <cr>=2):
Input file name: class.r
Input rank of top variable to split root node ([1:431], <cr>=1):
Run GUIDE with the command: guide < class.in

5.1.2 Contents of output file

Classification tree
Pruning by cross-validation
Data description file: ce2021class.dsc
Training sample file: ce2021.txt
Missing value code: NA
Records in data file start on line 2
Number of M variables associated with C variables: 19
384 N variables changed to S
D variable is INTRDVX_
Number of records in data file: 3965
Length of longest entry in data file: 11
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Number of classes: 3
Warning: S variable DIRACC is constant
Warning: S variable TOTHVHRP is constant
Warning: S variable TOTHVHRC is constant
Warning: S variable ROTHRFLEC is constant
Warning: S variable WELFREBX is constant
Smallest and largest positive weights are 1.0725E+03 and 9.3902E+04
Training sample class proportions of D variable INTRDVX_:

<table>
<thead>
<tr>
<th>Class</th>
<th>#Cases</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>1478</td>
<td>0.37276166</td>
</tr>
<tr>
<td>D</td>
<td>2431</td>
<td>0.61311475</td>
</tr>
<tr>
<td>T</td>
<td>56</td>
<td>0.01412358</td>
</tr>
</tbody>
</table>
5.1 Classification tree 5 MISSING-VALUE FLAG VARIABLES: CE DATA

Summary information for training sample of size 3965

- d=dependent, b=split and fit cat variable using indicator variables,
- c=split-only categorical, i=fit-only categorical (via indicators),
- s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
- m=missing-value flag variable, p=periodic variable, w=weight

Levels of M variables are for missing values in associated variables

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DIRACC</td>
<td>s</td>
<td>1.000</td>
<td>1.000</td>
<td>170</td>
</tr>
<tr>
<td>2</td>
<td>DIRACC_</td>
<td>m</td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>AGE_REF</td>
<td>s</td>
<td>18.00</td>
<td>87.00</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>AGE_REF_</td>
<td>m</td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>AGE2</td>
<td>s</td>
<td>21.00</td>
<td>87.00</td>
<td>1734</td>
</tr>
<tr>
<td>6</td>
<td>AGE2_</td>
<td>m</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>INTRDVX</td>
<td>d</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>407</td>
<td>IRAB</td>
<td>s</td>
<td>1.000</td>
<td>6.000</td>
<td>3831</td>
</tr>
<tr>
<td>409</td>
<td>IRAB_</td>
<td>m</td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>WHLFYR</td>
<td>c</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>547</td>
<td>WHLFYR_</td>
<td>m</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>549</td>
<td>FFTAXOWE</td>
<td>s</td>
<td>-0.3368E+05</td>
<td>0.3997E+06</td>
<td></td>
</tr>
<tr>
<td>550</td>
<td>FSTAXOWE</td>
<td>s</td>
<td>-3309</td>
<td>0.7223E+05</td>
<td></td>
</tr>
</tbody>
</table>

Total #cases w/ #missing

<table>
<thead>
<tr>
<th>#cases</th>
<th>miss. D</th>
<th>ord. vals</th>
<th>#X-var</th>
<th>#N-var</th>
<th>#F-var</th>
<th>#S-var</th>
</tr>
</thead>
<tbody>
<tr>
<td>3965</td>
<td>0</td>
<td>3965</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>384</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#P-var</th>
<th>#M-var</th>
<th>#B-var</th>
<th>#C-var</th>
<th>#I-var</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>116</td>
<td>0</td>
<td>47</td>
<td>0</td>
</tr>
</tbody>
</table>

Number of cases used for training: 3965
Number of split variables: 431
Number of cases excluded due to 0 W or missing D variable: 0

Constant fitted to cases with missing values in regressor variables
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE’s for pruned tree: 0.2500

Warning: No interaction tests; too many predictor variables
Simple node models
Estimated priors
Unit misclassification costs
Warning: All positive weights treated as 1
Univariate split highest priority
No interaction splits
No linear splits
Split values for N and S variables based on exhaustive search
Maximum number of split levels: 13
Minimum node sample size: 39
Top-ranked variables and 1-df chi-squared values at root node
1 0.2367E+03 INCLASS2
2 0.1936E+03 STATE
3 0.1519E+03 ERANKH
4 0.1350E+03 PSU
:
379 0.1031E-01 PUBTRACQ
380 0.2768E-02 CARTKUPQ
381 0.2626E-02 OTHLOAN
382 0.6293E-03 TEXTILCQ

Size and CV mean cost and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Nodes</th>
<th>Mean Cost</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median Cost</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>74</td>
<td>3.190E-01</td>
<td>7.402E-03</td>
<td>6.837E-03</td>
<td>3.161E-01</td>
<td>1.242E-02</td>
</tr>
<tr>
<td>2</td>
<td>73</td>
<td>3.190E-01</td>
<td>7.402E-03</td>
<td>6.837E-03</td>
<td>3.161E-01</td>
<td>1.242E-02</td>
</tr>
<tr>
<td>3</td>
<td>71</td>
<td>3.190E-01</td>
<td>7.402E-03</td>
<td>6.837E-03</td>
<td>3.161E-01</td>
<td>1.242E-02</td>
</tr>
<tr>
<td>:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29++</td>
<td>28</td>
<td>3.120E-01</td>
<td>7.358E-03</td>
<td>6.105E-03</td>
<td>3.052E-01</td>
<td>1.040E-02</td>
</tr>
<tr>
<td>30</td>
<td>27</td>
<td>3.107E-01</td>
<td>7.350E-03</td>
<td>5.616E-03</td>
<td>3.086E-01</td>
<td>7.575E-03</td>
</tr>
<tr>
<td>31</td>
<td>25</td>
<td>3.105E-01</td>
<td>7.348E-03</td>
<td>5.663E-03</td>
<td>3.086E-01</td>
<td>7.916E-03</td>
</tr>
<tr>
<td>32**</td>
<td>21</td>
<td>3.072E-01</td>
<td>7.326E-03</td>
<td>4.146E-03</td>
<td>3.086E-01</td>
<td>5.017E-03</td>
</tr>
<tr>
<td>33</td>
<td>14</td>
<td>3.115E-01</td>
<td>7.354E-03</td>
<td>6.153E-03</td>
<td>3.174E-01</td>
<td>1.043E-02</td>
</tr>
<tr>
<td>34</td>
<td>11</td>
<td>3.102E-01</td>
<td>7.346E-03</td>
<td>7.847E-03</td>
<td>3.115E-01</td>
<td>1.483E-02</td>
</tr>
<tr>
<td>35</td>
<td>8</td>
<td>3.185E-01</td>
<td>7.399E-03</td>
<td>6.886E-03</td>
<td>3.266E-01</td>
<td>1.041E-02</td>
</tr>
<tr>
<td>37</td>
<td>2</td>
<td>3.417E-01</td>
<td>7.532E-03</td>
<td>1.010E-02</td>
<td>3.317E-01</td>
<td>1.779E-02</td>
</tr>
<tr>
<td>38</td>
<td>1</td>
<td>3.869E-01</td>
<td>7.735E-03</td>
<td>8.543E-03</td>
<td>3.851E-01</td>
<td>1.446E-02</td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 21 terminal nodes
0-SE tree based on median is marked with + and has 28 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
** tree same as -- tree
+ tree same as ++ tree
* tree same as ** tree
* tree same as -- tree

Following tree is based on mean CV with naive SE estimate (**)
Structure of final tree. Each terminal node is marked with a T.
## 5.1 Classification tree

5. MISS-VALUE FLAG VARIABLES: CE DATA

Node cost is node misclassification cost divided by number of training cases

<table>
<thead>
<tr>
<th>Node label</th>
<th>Total cases</th>
<th>Train cases</th>
<th>Predicted class</th>
<th>Node cost</th>
<th>Node split</th>
<th>Interacting variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3965</td>
<td>3965</td>
<td>D</td>
<td>3.869E-01</td>
<td>INCLASS2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3717</td>
<td>3717</td>
<td>D</td>
<td>3.613E-01</td>
<td>PSU</td>
<td></td>
</tr>
<tr>
<td>4T</td>
<td>126</td>
<td>126</td>
<td>C</td>
<td>1.668E-01</td>
<td>OCCUCOD2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>3591</td>
<td>3591</td>
<td>D</td>
<td>3.448E-01</td>
<td>STATE</td>
<td></td>
</tr>
<tr>
<td>10T</td>
<td>1360</td>
<td>1360</td>
<td>D</td>
<td>2.154E-01</td>
<td>INCNONW2</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>2231</td>
<td>2231</td>
<td>D</td>
<td>4.236E-01</td>
<td>RETSURVX</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>1609</td>
<td>1609</td>
<td>D</td>
<td>4.201E-01</td>
<td>FINDRETX</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>606</td>
<td>968</td>
<td>D</td>
<td>4.926E-01</td>
<td>INCLASS2</td>
<td></td>
</tr>
<tr>
<td>88</td>
<td>538</td>
<td>538</td>
<td>D</td>
<td>4.108E-01</td>
<td>STOCKX</td>
<td></td>
</tr>
<tr>
<td>176</td>
<td>483</td>
<td>483</td>
<td>D</td>
<td>4.161E-01</td>
<td>RENTEQVX</td>
<td></td>
</tr>
<tr>
<td>352</td>
<td>358</td>
<td>358</td>
<td>D</td>
<td>3.520E-01</td>
<td>STATE</td>
<td></td>
</tr>
<tr>
<td>704T</td>
<td>45</td>
<td>45</td>
<td>C</td>
<td>2.668E-01</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>705T</td>
<td>313</td>
<td>313</td>
<td>D</td>
<td>2.971E-01</td>
<td>INCNONW2</td>
<td></td>
</tr>
<tr>
<td>353T</td>
<td>125</td>
<td>125</td>
<td>C</td>
<td>4.001E-01</td>
<td>ROOMSQ</td>
<td></td>
</tr>
<tr>
<td>177T</td>
<td>55</td>
<td>55</td>
<td>D</td>
<td>3.636E-01</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>89</td>
<td>430</td>
<td>430</td>
<td>C</td>
<td>4.395E-01</td>
<td>EARNCOMP</td>
<td></td>
</tr>
<tr>
<td>178T</td>
<td>47</td>
<td>47</td>
<td>C</td>
<td>6.170E-01</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>179</td>
<td>383</td>
<td>383</td>
<td>C</td>
<td>4.178E-01</td>
<td>REFGEN</td>
<td></td>
</tr>
<tr>
<td>358</td>
<td>296</td>
<td>296</td>
<td>C</td>
<td>3.818E-01</td>
<td>REFGEN</td>
<td></td>
</tr>
<tr>
<td>716</td>
<td>148</td>
<td>148</td>
<td>C</td>
<td>4.122E-01</td>
<td>DMSRVCQ</td>
<td></td>
</tr>
<tr>
<td>1432</td>
<td>107</td>
<td>107</td>
<td>C</td>
<td>3.272E-01</td>
<td>STATE</td>
<td></td>
</tr>
<tr>
<td>2864T</td>
<td>41</td>
<td>41</td>
<td>D</td>
<td>3.659E-01</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2865T</td>
<td>66</td>
<td>66</td>
<td>C</td>
<td>1.365E-01</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>1433T</td>
<td>41</td>
<td>41</td>
<td>D</td>
<td>4.146E-01</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>717</td>
<td>148</td>
<td>148</td>
<td>C</td>
<td>3.514E-01</td>
<td>CASHCOCQ</td>
<td></td>
</tr>
<tr>
<td>1434T</td>
<td>109</td>
<td>109</td>
<td>C</td>
<td>2.661E-01</td>
<td>STATE</td>
<td></td>
</tr>
<tr>
<td>1435T</td>
<td>39</td>
<td>39</td>
<td>D</td>
<td>4.103E-01</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>359</td>
<td>87</td>
<td>87</td>
<td>D</td>
<td>4.598E-01</td>
<td>PROPTXPQ</td>
<td></td>
</tr>
<tr>
<td>718T</td>
<td>45</td>
<td>45</td>
<td>D</td>
<td>3.333E-01</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>719T</td>
<td>42</td>
<td>42</td>
<td>C</td>
<td>4.048E-01</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>641</td>
<td>641</td>
<td>D</td>
<td>3.105E-01</td>
<td>HLFBATHQ</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>350</td>
<td>350</td>
<td>D</td>
<td>3.514E-01</td>
<td>FSTAXOBE</td>
<td></td>
</tr>
<tr>
<td>180</td>
<td>299</td>
<td>299</td>
<td>D</td>
<td>2.977E-01</td>
<td>OCCUCOD1</td>
<td></td>
</tr>
<tr>
<td>360T</td>
<td>43</td>
<td>43</td>
<td>C</td>
<td>3.489E-01</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>361T</td>
<td>256</td>
<td>256</td>
<td>D</td>
<td>2.383E-01</td>
<td>ELCTRCPQ</td>
<td></td>
</tr>
<tr>
<td>181T</td>
<td>51</td>
<td>51</td>
<td>C</td>
<td>3.334E-01</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>91T</td>
<td>291</td>
<td>291</td>
<td>D</td>
<td>2.612E-01</td>
<td>HOUSOPPPQ</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>622</td>
<td>622</td>
<td>D</td>
<td>4.325E-01</td>
<td>RETSURVX</td>
<td></td>
</tr>
<tr>
<td>46T</td>
<td>86</td>
<td>86</td>
<td>C</td>
<td>1.397E-01</td>
<td>STATE</td>
<td></td>
</tr>
<tr>
<td>47T</td>
<td>536</td>
<td>536</td>
<td>D</td>
<td>3.638E-01</td>
<td>STOCKYRX</td>
<td></td>
</tr>
<tr>
<td>3T</td>
<td>248</td>
<td>248</td>
<td>C</td>
<td>2.299E-01</td>
<td>FINCBTAX</td>
<td></td>
</tr>
</tbody>
</table>

Warning: tree very large, omitting node numbers in LaTeX file
Number of terminal nodes of final tree: 21
Total number of nodes of final tree: 41
Second best split variable (based on curvature test) at root node is STATE

Classification tree:
For categorical variable splits, values not in training data go to the right

Node 1: INCLASS2 <= 6.5000000
  Node 2: PSU = "S49F"
    Node 4: C
  Node 2: PSU /= "S49F"
            "41", "47", "48", "49"
      Node 10: D
            "41", "47", "48", "49"
      Node 11: RETSURVX = NA & RETS_RVX = "A"
  Node 22: FINDRETX <= 391.50000
    Node 44: INCLASS2 <= 4.5000000
      Node 88: STOCKX = NA & STOCKX_ = "A"
        Node 176: RENTEQVX <= 1990.5000 or NA
          Node 352: STATE = "13", "17", "19", "22", "28", "32", "45", "46"
             Node 704: C
             Node 705: D
        Node 176: RENTEQVX > 1990.5000
          Node 353: C
          Node 88: not (STOCKX = NA & STOCKX_ = "A")
          Node 177: D
    Node 44: INCLASS2 > 4.5000000 or NA
      Node 89: EARNCOMP = "7", "8"
        Node 178: C
        Node 89: EARNCOMP /= "7", "8"
          Node 179: REFGEN <= 4.5000000
            Node 358: REFGEN <= 3.5000000
              Node 716: DOMSRVCQ <= 12.000000
                Node 1432: STATE = "26", "39", "42", "46", "51", "NA"
                  Node 2864: D
                Node 1432: STATE /= "26", "39", "42", "46", "51", "NA"
                  Node 2865: C
              Node 716: DOMSRVCQ > 12.000000 or NA
                Node 1433: D
            Node 358: REFGEN > 3.5000000 or NA
              Node 717: CASHCDCQ <= 30.000000
                Node 1434: C
Node 717: CASHCOCQ > 30.000000 or NA
Node 1435: D
Node 179: REFGEN > 4.5000000 or NA
Node 359: PROPTXPQ <= 287.50000
  Node 718: D
  Node 359: PROPTXPQ > 287.50000 or NA
  Node 719: C
Node 22: FINDRETX > 391.500000 or NA
Node 45: HLFBATHQ <= 0.50000000
  Node 90: FSTAXOWE <= 10256.500000
    Node 180: OCCUCOD1 = "6", "7", "11", "12", "15"
      Node 360: C
    Node 180: OCCUCOD1 /= "6", "7", "11", "12", "15"
      Node 361: D
    Node 90: FSTAXOWE > 10256.50000 or NA
    Node 181: C
    Node 45: HLFBATHQ > 0.50000000 or NA
    Node 91: D
Node 11: not (RETSURVX = NA & RETS_RVX = "A")
  Node 23: RETSURVX = NA
    Node 46: C
    Node 23: RETSURVX /= NA
    Node 47: D
Node 1: INCLASS2 > 6.5000000 or NA
  Node 3: C

***********************************************************************
Predictor means below are weighted means of cases with no missing values.

Node 1: Intermediate node
A case goes into Node 2 if INCLASS2 <= 6.5000000
INCLASS2 mean = 4.6266367
Class    Number  Posterior
  C      1478  0.3728E+00
  D      2431  0.6131E+00
  T       56  0.1412E-01
Number of training cases misclassified = 1534
Predicted class is D

Node 2: Intermediate node
A case goes into Node 4 if PSU = "S49F"
PSU mode = "NA"
Class    Number  Posterior
  C       1287  0.3462E+00
  D       2374  0.6387E+00
  T        56  0.1507E-01

Wei-Yin Loh 80 GUIDE manual
5.1 Classification tree

Number of training cases misclassified = 1343
Predicted class is D

Node 4: Terminal node

<table>
<thead>
<tr>
<th>Class</th>
<th>Number</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>105</td>
<td>0.8332E+00</td>
</tr>
<tr>
<td>D</td>
<td>21</td>
<td>0.1668E+00</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0.3561E-05</td>
</tr>
</tbody>
</table>

Number of training cases misclassified = 21
Predicted class is C

Node 47: Terminal node

<table>
<thead>
<tr>
<th>Class</th>
<th>Number</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>185</td>
<td>0.3451E+00</td>
</tr>
<tr>
<td>D</td>
<td>341</td>
<td>0.6362E+00</td>
</tr>
<tr>
<td>T</td>
<td>10</td>
<td>0.1866E-01</td>
</tr>
</tbody>
</table>

Number of training cases misclassified = 195
Predicted class is D

Node 3: Terminal node

<table>
<thead>
<tr>
<th>Class</th>
<th>Number</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>191</td>
<td>0.7701E+00</td>
</tr>
<tr>
<td>D</td>
<td>57</td>
<td>0.2299E+00</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0.3561E-05</td>
</tr>
</tbody>
</table>

Number of training cases misclassified = 57
Predicted class is C

Classification matrix for training sample:

<table>
<thead>
<tr>
<th>Predicted class</th>
<th>True class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
</tr>
<tr>
<td>C</td>
<td>720</td>
</tr>
<tr>
<td>D</td>
<td>758</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>1478</td>
</tr>
</tbody>
</table>

Number of cases used for tree construction: 3965
Number misclassified: 1069
Resubstitution estimate of mean misclassification cost: 0.26960908

Observed and fitted values are stored in class.fit
LaTeX code for tree is in class.tex
R code is stored in class.r
Figure 8 shows the classification tree. Three different kinds of splits on missing values are exhibited in these intermediate nodes:

**Node 1:** $\text{INCLASS2} \leq 6.50$ with missing values going right.

**Node 11:** $\text{RETS} \_\text{RVX} = A$ (some missing values go left, all non-missing and some missing values go right).

**Node 23:** $\text{RET} \_\text{SURVX} = \text{NA}$ (all missing values go left, nonmissing go right).

Owing to the small number of cases of $\text{INTRDVX} _\text{T}$, the tree has no terminal node that predicts this class. The top several lines of the file of fitted values `class.fit` are given below. The posterior probabilities of predicting class T are very low (see Section 4.1.4 for the calculation of the posterior probabilities).

<table>
<thead>
<tr>
<th>train</th>
<th>node</th>
<th>observed</th>
<th>predicted</th>
<th>&quot;P(C)&quot;</th>
<th>&quot;P(D)&quot;</th>
<th>&quot;P(T)&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>4</td>
<td>&quot;C&quot;</td>
<td>&quot;C&quot;</td>
<td>0.83322E+00</td>
<td>0.16678E+00</td>
<td>0.35612E-05</td>
</tr>
<tr>
<td>y</td>
<td>10</td>
<td>&quot;D&quot;</td>
<td>&quot;D&quot;</td>
<td>0.19779E+00</td>
<td>0.78456E+00</td>
<td>0.17647E-01</td>
</tr>
<tr>
<td>y</td>
<td>10</td>
<td>&quot;D&quot;</td>
<td>&quot;D&quot;</td>
<td>0.19779E+00</td>
<td>0.78456E+00</td>
<td>0.17647E-01</td>
</tr>
<tr>
<td>y</td>
<td>1434</td>
<td>&quot;D&quot;</td>
<td>&quot;C&quot;</td>
<td>0.73385E+00</td>
<td>0.26614E+00</td>
<td>0.35612E-05</td>
</tr>
<tr>
<td>y</td>
<td>10</td>
<td>&quot;D&quot;</td>
<td>&quot;D&quot;</td>
<td>0.19779E+00</td>
<td>0.78456E+00</td>
<td>0.17647E-01</td>
</tr>
<tr>
<td>y</td>
<td>91</td>
<td>&quot;T&quot;</td>
<td>&quot;D&quot;</td>
<td>0.23711E+00</td>
<td>0.73883E+00</td>
<td>0.24055E-01</td>
</tr>
<tr>
<td>y</td>
<td>91</td>
<td>&quot;D&quot;</td>
<td>&quot;D&quot;</td>
<td>0.23711E+00</td>
<td>0.73883E+00</td>
<td>0.24055E-01</td>
</tr>
<tr>
<td>y</td>
<td>10</td>
<td>&quot;D&quot;</td>
<td>&quot;D&quot;</td>
<td>0.19779E+00</td>
<td>0.78456E+00</td>
<td>0.17647E-01</td>
</tr>
</tbody>
</table>
Figure 8: GUIDE classification tree for predicting INTRDVX using estimated priors and unit misclassification costs. At each split, an observation goes to the left branch if and only if the condition is satisfied. Symbol ‘≤’ stands for ‘≤ or missing’. \( S_1 = \{2, 6, 10, 11, 21, 24, 25, 27, 31, 40, 41, 47, 48, 49\} \). \( S_2 = \{13, 17, 19, 22, 28, 32, 45, 46\} \). \( S_3 = \{7, 8\} \). \( S_4 = \{26, 39, 42, 46, 51, NA\} \). \( S_5 = \{6, 7, 11, 12, 15\} \). Predicted classes and sample sizes (in italics) printed below terminal nodes; class sample proportions for \( \text{INTRDVX}_- = \text{C}, \text{D}, \text{T} \), respectively, beside nodes. Second best split variable at root node is \( \text{STATE} \).
6 Least squares regression: CE data

GUIDE can fit least-squares (LS), quantile, Poisson, proportional hazards, and least-
median-of-squares (LMS) regression tree models. We illustrate least squares and
quantile models with the CE data, using INTRDVX as the dependent (d) variable and
excluding (x) its flag INTRDVX_. The description file is ce2021reg.dsc, which sets
FINLWT21 as a weight (w) variable.

6.1 Piecewise constant

6.1.1 Input file creation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: cons.in
   Input 1 for model fitting, 2 for importance or DIF scoring,
     3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: cons.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1): 2
   Choose type of regression model:
     1=linear, 2=quantile, 3=Poisson, 4=censored response,
     5=multiresponse or itemresponse, 6=longitudinal data (with T variables),
     7=binary logistic regression.
   Input choice ([1:7], <cr>=1):
     Input 1 for least squares, 2 least median of squares ([1:2], <cr>=1):
     Choose complexity of model to use at each node:
     Choose 0 for stepwise linear regression (recommended for prediction)
     Choose 1 for multiple regression
     Choose 2 for best simple polynomial in one N or F variable
     Choose 3 for constant fit (recommended for interpretability or if there is an R variable)
     0: stepwise linear, 1: multiple linear, 2: best simple polynomial, 3: constant,
     4: best simple stepwise ANCOVA ([0:4], <cr>=3):
     Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
     Input name of data description file (max 100 characters);
     enclose with matching quotes if it has spaces: ce2021reg.dsc
Reading data description file ...  
Training sample file: ce2021.txt
Missing value code: NA
Records in data file start on line 2
Number of M variables associated with C variables: 19
384 N variables changed to S
D variable is INTRDVX
Reading data file ...
Number of records in data file: 3965
Length of longest entry in data file: 11
Checking for missing values ...
Finished checking
Missing values found in D variable
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Finding number of levels of M variables associated with C variables ...
Assigning integer codes to values of 47 categorical variables
Finished assigning codes to 10 categorical variables
Finished assigning codes to 20 categorical variables
Finished assigning codes to 30 categorical variables
Finished assigning codes to 40 categorical variables
Associating missing values of N, P and S variables with M variable codes ...
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...
Data checks complete
Creating missing value indicators ...
Rereading data ...
Warning: S variable DIRACC is constant
Warning: S variable TOTHVHRP is constant
Warning: S variable TOTHVHRC is constant
Warning: S variable ROTHRFC is constant
Warning: S variable WELFREBX is constant
Warning: S variable OTHLYRBX is constant
Warning: S variable OTHLNYRB is constant
Smallest positive weight: 1.0725E+03
Largest positive weight: 9.3902E+04

<table>
<thead>
<tr>
<th>Total #cases w/</th>
<th>#cases miss. D ord. vals</th>
<th>#X-var</th>
<th>#N-var</th>
<th>#F-var</th>
<th>#S-var</th>
</tr>
</thead>
<tbody>
<tr>
<td>3965</td>
<td>1478</td>
<td>3965</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>384</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Weight variable FINLWT21 in column: 31
Number of cases used for training: 2487
Number of split variables: 431
Number of cases excluded due to 0 W or missing D variable: 1478
Finished reading data file
Input 1 for unweighted, 2 for weighted error estimates during pruning ([1:2], <cr>=2):
Warning: No interaction tests; too many predictor variables
Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):
Input file name to store LaTeX code (use .tex as suffix): cons.tex
You can store the variables and/or values used to split and fit in a file
6.1 Piecewise constant

Choose 1 to skip this step, 2 to store split and fit variables, 3 to store split variables and their values
Input your choice ([1:3], <cr>=1):
Input 2 to save fitted values and node IDs, 1 otherwise ([1:2], <cr>=2):
Input name of file to store node ID and fitted value of each case: cons.fit
Input 2 to write R function for predicting new cases, 1 otherwise ([1:2], <cr>=2):
Input file name: cons.r
Input rank of top variable to split root node ([1:431], <cr>=1):
Input file is created!
Run GUIDE with the command: guide < cons.in

6.1.2 Contents of cons.out

Least squares regression tree
Pruning by cross-validation
Data description file: ce2021reg.dsc
Training sample file: ce2021.txt
Missing value code: NA
Records in data file start on line 2
Number of M variables associated with C variables: 19
384 N variables changed to S
D variable is INTRDVX
Piecewise constant model
Number of records in data file: 3965
Length of longest entry in data file: 11
Missing values found in D variable
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Warning: S variable DIRACC is constant
Warning: S variable TOTHVHRP is constant
Warning: S variable TOTHVHRC is constant
Warning: S variable ROTHRFCLC is constant
Warning: S variable WELFREBX is constant
Warning: S variable OTHLYRBX is constant
Warning: S variable OTHLNYRB is constant
Smallest and largest positive weights are 1.0725E+03 and 9.3902E+04

Summary information for training sample of size 2487 (excluding observations with non-positive weight or missing values in d, e, t, r or z variables)
d=dependent, b=split and fit cat variable using indicator variables, c=split-only categorical, i=fit-only categorical (via indicators), s=split-only numerical, n=split and fit numerical, f=fit-only numerical, m=missing-value flag variable, p=periodic variable, w=weight
Levels of M variables are for missing values in associated variables

Wei-Yin Loh 86 GUIDE manual
### 6.1 Piecewise constant LEAST SQUARES REGRESSION: CE DATA

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 DIRACC</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td>125</td>
</tr>
<tr>
<td>2 DIRACC_</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>3 AGE_REF</td>
<td>19.00</td>
<td>87.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 AGE_REF_</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>5 AGE2</td>
<td>21.00</td>
<td>87.00</td>
<td></td>
<td>1092</td>
</tr>
<tr>
<td>6 AGE2_</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>30 FIND_ETX</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>31 FINLWT21</td>
<td>1072.</td>
<td>0.939E+05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32 FJSSDEDX</td>
<td>0.000</td>
<td>0.4366E+05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33 FJSS_EDX</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>406 INTRDVX</td>
<td>1.000</td>
<td>0.1413E+06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>547 WHLFYR</td>
<td></td>
<td></td>
<td></td>
<td>1 2487</td>
</tr>
<tr>
<td>548 WHLFYR_</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>549 FFTAXOWE</td>
<td>-0.3368E+05</td>
<td>0.3380E+06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>550 FSTAXOWE</td>
<td>-3074.</td>
<td>0.5654E+05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total #cases w/ missing

- #cases: 3965
- miss. D ord. vals: 1478
- #X-var: 1
- #N-var: 0
- #F-var: 0
- #S-var: 384
- #P-var: 0
- #M-var: 1
- #B-var: 47
- #C-var: 0
- #I-var: 0

Weight variable FINLWT21 in column: 31
Number of cases used for training: 2487
Number of split variables: 431
Number of cases excluded due to 0 W or missing D variable: 1478

Constant fitted to cases with missing values in regressor variables
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE’s for pruned tree: 0.2500

Weighted error estimates used for pruning
Warning: No interaction tests; too many predictor variables
No nodewise interaction tests
Split values for N and S variables based on exhaustive search
Maximum number of split levels: 12
Minimum node sample size: 24
Top-ranked variables and 1-df chi-squared values at root node
1 0.8297E+02  REFGEN
2 0.8111E+02  AGE_REF

Wei-Yin Loh 87 GUIDE manual
6.1 Piecewise constant LEAST SQUARES REGRESSION: CE DATA

3 0.7066E+02 INCNONW1
4 0.6985E+02 STOCKX

Size and CV MSE and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Tnodes</th>
<th>Mean MSE</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median MSE</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>79</td>
<td>5.890E+12</td>
<td>7.198E+11</td>
<td>4.740E+11</td>
<td>5.838E+12</td>
<td>7.710E+11</td>
</tr>
<tr>
<td>2</td>
<td>78</td>
<td>5.890E+12</td>
<td>7.198E+11</td>
<td>4.740E+11</td>
<td>5.838E+12</td>
<td>7.710E+11</td>
</tr>
<tr>
<td>39</td>
<td>21</td>
<td>5.860E+12</td>
<td>7.167E+11</td>
<td>4.808E+11</td>
<td>5.830E+12</td>
<td>8.050E+11</td>
</tr>
<tr>
<td>40+</td>
<td>14</td>
<td>5.859E+12</td>
<td>7.168E+11</td>
<td>4.806E+11</td>
<td>5.822E+12</td>
<td>8.039E+11</td>
</tr>
<tr>
<td>41</td>
<td>9</td>
<td>5.869E+12</td>
<td>7.194E+11</td>
<td>4.902E+11</td>
<td>5.849E+12</td>
<td>8.075E+11</td>
</tr>
<tr>
<td>42**</td>
<td>8</td>
<td>5.789E+12</td>
<td>7.184E+11</td>
<td>5.133E+11</td>
<td>5.860E+12</td>
<td>8.042E+11</td>
</tr>
<tr>
<td>45</td>
<td>1</td>
<td>8.287E+12</td>
<td>1.032E+12</td>
<td>6.955E+11</td>
<td>8.542E+12</td>
<td>7.418E+11</td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 8 terminal nodes
0-SE tree based on median is marked with + and has 14 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on median and bootstrap SE is marked with ++
** tree same as ++ tree
** tree same as -- tree
++ tree same as -- tree
* tree same as ** tree
* tree same as ++ tree
* tree same as -- tree

Following tree is based on mean CV with naive SE estimate (**)

Structure of final tree. Each terminal node is marked with a T.

D-mean is weighted mean of INTRDVX in the node
Cases fit give the number of cases used to fit node
MSE is residual sum of squares divided by number of cases in node

<table>
<thead>
<tr>
<th>Node</th>
<th>Total Cases</th>
<th>Matrix Node</th>
<th>Node Split</th>
<th>D-mean</th>
<th>MSE variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>label</td>
<td>cases</td>
<td>fit rank</td>
<td>rank</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2487</td>
<td>2487</td>
<td>1</td>
<td>5.131E+03</td>
<td>8.287E+12</td>
</tr>
<tr>
<td>2</td>
<td>1914</td>
<td>1914</td>
<td>1</td>
<td>6.476E+03</td>
<td>1.027E+13</td>
</tr>
<tr>
<td>4T</td>
<td>1345</td>
<td>1345</td>
<td>1</td>
<td>2.823E+03</td>
<td>1.657E+12</td>
</tr>
<tr>
<td>5</td>
<td>569</td>
<td>569</td>
<td>1</td>
<td>1.498E+04</td>
<td>2.828E+13</td>
</tr>
<tr>
<td>10</td>
<td>75</td>
<td>75</td>
<td>1</td>
<td>5.278E+04</td>
<td>7.423E+13</td>
</tr>
</tbody>
</table>
6.1 Piecewise constant 6 LEAST SQUARES REGRESSION: CE DATA

Number of terminal nodes of final tree: 8  
Total number of nodes of final tree: 15

Second best split variable (based on curvature test) at root node is AGE_REF

Regression tree:
For categorical variable splits, values not in training data go to the right

Node 1: REFGEN <= 4.5000000
 Node 2: INC_RANK <= 0.84018625
    Node 4: INTRDVX-mean = 2822.6445
 Node 2: INC_RANK > 0.84018625 or NA
    Node 5: EARNCOMP = "8"
       Node 10: RETSURV = "1"
          Node 20: INTRDVX-mean = 27641.282
          Node 10: RETSURV /= "1"
             Node 21: INTRDVX-mean = 85859.276
    Node 5: EARNCOMP /= "8"
    Node 11: FFTAXOWE <= 27769.500
        Node 22: INTRDVX-mean = 2646.5367
        Node 11: FFTAXOWE > 27769.500 or NA
            Node 23: AGE2 <= 56.500000 or NA
                Node 46: INTRDVX-mean = 8036.3341
                Node 23: AGE2 > 56.500000
                   Node 47: BATHRMQ <= 2.5000000
                      Node 94: INTRDVX-mean = 10866.520
                      Node 47: BATHRMQ > 2.5000000 or NA
                         Node 95: INTRDVX-mean = 46702.398
    Node 1: REFGEN > 4.5000000 or NA
        Node 3: INTRDVX-mean = 980.35292

***********************************************************************
Predictor means below are weighted means of cases with no missing values.

WARNING: p-values below not adjusted for split search. For a bootstrap solution see:

Wei-Yin Loh
6.1 Piecewise constant  


Node 1: Intermediate node
A case goes into Node 2 if REFGEN <= 4.500000
REFGEN mean = 3.4302980

Coefficients of least squares regression function:
Regressor Coefficient t-stat p-value
Constant 5131. 12.43 0.2220E-15
INTRDVX mean = 5130.60

Node 2: Intermediate node
A case goes into Node 4 if INC_RANK <= 0.84018625
INC_RANK mean = 0.64479638

Coefficients of least squares regression functions:
Regressor Coefficient t-stat p-value
Constant 2823. 11.12 0.3331E-15
INTRDVX mean = 2822.64

Node 4: Terminal node

Coefficients of least squares regression functions:
Regressor Coefficient t-stat p-value
Constant 0.1087E+05 2.340 0.2368E-01
INTRDVX mean = 10866.5

Node 94: Terminal node

Coefficients of least squares regression functions:
Regressor Coefficient t-stat p-value
Constant 0.4670E+05 6.082 0.2775E-06
INTRDVX mean = 46702.4

Node 3: Terminal node
6.2 Piecewise simple polynomial

GUIDE can also fit a simple polynomial regression model in each node of the form

\[ y = \beta_0 + \sum_{k=1}^{p} \beta_k x^k + \epsilon \]  

(1)

where \( p \) is the degree of polynomial desired and \( x \) is selected from the set of \( n \) and \( f \) variables. The variable \( x \) is the one among all \( n \) and \( f \) variables that yields the smallest sum of squared residuals. Variable \( x \) can vary from node to node. If there are missing values in the \( x \) variable, GUIDE fits two separate models to the data in the node: model (1) to the observations with complete values in \( x \) and \( y \) and a
Figure 9: GUIDE v.41.0 0.250-SE piecewise-constant weighted least-squares regression tree for predicting INTRDVX. At each split, an observation goes to the left branch if and only if the condition is satisfied. Symbol ‘≤∗’ stands for ‘≤ or missing’. Sample size (in italics) and weighted mean of INTRDVX printed below nodes. Terminal nodes with means above and below value of 5130.6 at root node are colored yellow and skyblue respectively. Second best split variable at root node is AGE_REF.
constant \((y = \beta_0 + \epsilon)\) to those with missing values in \(x\). This is equivalent to imputing missing \(x\) values with a constant \(c\) and adding the missing-value indicator \(I(x = NA)\) as linear predictor:

\[
y = \beta_0 + \sum_{k=1}^{p} \beta_k x_1^k + \beta_2 x_2 + \epsilon
\]

where \(x_1 = xI(x \neq NA) + cI(x = NA)\) and \(x_2 = I(x = NA)\). The predicted values are independent of \(c\) but the least-squares estimates of the \(\beta\) coefficients are not.

**Truncation note:** Extrapolation can adversely affect the prediction accuracy of parametric models. To guard against extrapolation, GUIDE has several options to truncate the predicted values, with the default being to truncate the predicted values if they fall outside the range of the observed values. The option of no truncation is available as well. Default truncation is used in this user guide.

### 6.2.1 Input file creation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   - Input your choice: 1
   - Name of batch input file: lin.in
   - Input 1 for model fitting, 2 for importance or DIF scoring,
     3 for data conversion ([1:3], \(<\text{cr}>=1\)):
   - Name of batch output file: lin.out
   - Input 1 for classification, 2 for regression, 3 for propensity score grouping
   - Input your choice ([1:3], \(<\text{cr}>=1\)):
     - Choose type of regression model:
       1=linear, 2=quantile, 3=Poisson, 4=censored response,
       5=multiproduct or itemresponse, 6=longitudinal data (with T variables),
       7=binary logistic regression.
     - Input choice ([1:7], \(<\text{cr}>=1\)):
     - Input 1 for least squares, 2 least median of squares ([1:2], \(<\text{cr}>=1\)):
     - Choose complexity of model to use at each node:
     - Choose 0 for stepwise linear regression (recommended for prediction)
     - Choose 1 for multiple regression
     - Choose 2 for best simple polynomial in one N or F variable
     - Choose 3 for constant fit (recommended for interpretability or if there is an R variable)
   0: stepwise linear, 1: multiple linear, 2: best simple polynomial, 3: constant,
   4: best simple stepwise ANCOVA ([0:4], \(<\text{cr}>=3\)):
   - Input 1 for default options, 2 otherwise ([1:2], \(<\text{cr}>=1\)):
   - Input name of data description file (max 100 characters);
     enclose with matching quotes if it has spaces: ce2021reg.dsc
   - Reading data description file ...
   - Training sample file: ce2021.txt
   - Missing value code: NA

Wei-Yin Loh 93 GUIDE manual
Records in data file start on line 2
Number of M variables associated with C variables: 19
D variable is INTRDVX
Reading data file ...
Number of records in data file: 3965
Length of longest entry in data file: 11
Checking for missing values ...
Finished checking
Missing values found in D variable
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Finding number of levels of M variables associated with C variables ...
Assigning integer codes to values of 47 categorical variables
Finished assigning codes to 10 categorical variables
Finished assigning codes to 20 categorical variables
Finished assigning codes to 30 categorical variables
Finished assigning codes to 40 categorical variables
Associating missing values of N, P and S variables with M variable codes ...
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...
Data checks complete
Creating missing value indicators ...
Rereading data ...
Warning: N variable DIRACC is constant
Warning: N variable TOTHVHRP is constant
Warning: N variable TOTHVHRC is constant
Warning: N variable ROTHRFLEC is constant
Warning: N variable WELFREBX is constant
Warning: N variable OTHLNYRBX is constant
Warning: N variable OTHLNYRBY is constant
Smallest positive weight: 1.0725E+03
Largest positive weight: 9.3902E+04
Total #cases w/ #missing
#cases miss. D ord. vals #X-var #N-var #F-var #S-var
3965 1478 3965 1 384 0 0
#P-var #M-var #B-var #C-var #I-var
0 116 0 47 0
Weight variable FINLWT21 in column: 31
Number of cases used for training: 2487
Number of split variables: 431
Number of cases excluded due to 0 W or missing D variable: 1478
Finished reading data file
Input 1 for unweighted, 2 for weighted error estimates during pruning ([1:2], <cr>=2):
Warning: No interaction tests; too many predictor variables
6.2 Piecewise simple polynomial LEAST SQUARES REGRESSION: CE DATA

Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):
Input file name to store LaTeX code (use .tex as suffix): lin.tex
You can store the variables and/or values used to split and fit in a file
Choose 1 to skip this step, 2 to store split and fit variables,
3 to store split variables and their values
Input your choice ([1:3], <cr>=1):
Input 2 to save regressor names in a file, 1 otherwise ([1:2], <cr>=2):
Input file name: lin.var
Input 2 to save fitted values and node IDs, 1 otherwise ([1:2], <cr>=2):
Input name of file to store node ID and fitted value of each case: lin.fit
Input 2 to write R function for predicting new cases, 1 otherwise ([1:2], <cr>=2):
Input file name: lin.r
Input rank of top variable to split root node ([1:431], <cr>=1):
Input file is created!
Run GUIDE with the command: guide < lin.in

6.2.2 Partial output

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Tnodes</th>
<th>Mean MSE</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median MSE</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>53</td>
<td>8.181E+12</td>
<td>1.005E+12</td>
<td>1.178E+12</td>
<td>7.611E+12</td>
<td>1.590E+12</td>
</tr>
<tr>
<td>2</td>
<td>52</td>
<td>8.181E+12</td>
<td>1.005E+12</td>
<td>1.178E+12</td>
<td>7.612E+12</td>
<td>1.590E+12</td>
</tr>
<tr>
<td>3</td>
<td>51</td>
<td>8.180E+12</td>
<td>1.005E+12</td>
<td>1.178E+12</td>
<td>7.612E+12</td>
<td>1.589E+12</td>
</tr>
<tr>
<td>:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27**</td>
<td>7</td>
<td>6.470E+12</td>
<td>7.870E+11</td>
<td>8.430E+11</td>
<td>5.256E+12</td>
<td>1.496E+12</td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 7 terminal nodes
0-SE tree based on median is marked with + and has 8 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
** tree same as ++ tree
** tree same as -- tree
++ tree same as -- tree
* tree same as ** tree
* tree same as ++ tree
* tree same as -- tree

Following tree is based on mean CV with naive SE estimate (**)
Structure of final tree. Each terminal node is marked with a T.

D-mean is weighted mean of INTRDVX in the node
Cases fit give the number of cases used to fit node
MSE and \( R^2 \) are based on all cases in node

<table>
<thead>
<tr>
<th>Node</th>
<th>Total</th>
<th>Cases</th>
<th>Matrix</th>
<th>Node</th>
<th>Node</th>
<th>Node</th>
<th>Node Split</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>label</td>
<td>cases</td>
<td>fit</td>
<td>rank</td>
<td>D-mean</td>
<td>MSE</td>
<td>( R^2 )</td>
<td>variable</td>
<td>variables</td>
</tr>
<tr>
<td>1</td>
<td>2487</td>
<td>209</td>
<td>2</td>
<td>5.131E+03</td>
<td>7.702E+12</td>
<td>0.0710</td>
<td>CUTENURE</td>
<td>+STOCKYRX</td>
</tr>
<tr>
<td>2</td>
<td>855</td>
<td>855</td>
<td>2</td>
<td>8.856E+03</td>
<td>1.219E+13</td>
<td>0.1292</td>
<td>FJSSDEDX</td>
<td>+FINCBTAX</td>
</tr>
<tr>
<td>4</td>
<td>578</td>
<td>578</td>
<td>2</td>
<td>9.919E+03</td>
<td>1.137E+13</td>
<td>0.2914</td>
<td>FINCBTAX</td>
<td>+FINCBTAX</td>
</tr>
<tr>
<td>8T</td>
<td>500</td>
<td>500</td>
<td>2</td>
<td>3.863E+03</td>
<td>1.323E+12</td>
<td>0.3043</td>
<td>INC_RANK</td>
<td>+ALCBEVCQ</td>
</tr>
<tr>
<td>9</td>
<td>78</td>
<td>47</td>
<td>2</td>
<td>5.014E+04</td>
<td>4.793E+13</td>
<td>0.2919</td>
<td>RETSURV</td>
<td>-RETSURVX</td>
</tr>
<tr>
<td>18T</td>
<td>47</td>
<td>47</td>
<td>2</td>
<td>2.613E+04</td>
<td>2.689E+13</td>
<td>0.2996</td>
<td>-</td>
<td>+FULOILCQ</td>
</tr>
<tr>
<td>19T</td>
<td>31</td>
<td>7</td>
<td>2</td>
<td>8.229E+04</td>
<td>3.579E+13</td>
<td>0.5077</td>
<td>-</td>
<td>-ROYESTX</td>
</tr>
<tr>
<td>5T</td>
<td>277</td>
<td>277</td>
<td>2</td>
<td>6.780E+03</td>
<td>4.930E+12</td>
<td>0.4882</td>
<td>PERINSCQ</td>
<td>+ETOTALC</td>
</tr>
<tr>
<td>3</td>
<td>1632</td>
<td>1087</td>
<td>2</td>
<td>3.221E+03</td>
<td>4.540E+12</td>
<td>0.1032</td>
<td>RENTEQVX</td>
<td>+RENTEQVX</td>
</tr>
<tr>
<td>6T</td>
<td>1558</td>
<td>1558</td>
<td>2</td>
<td>2.084E+03</td>
<td>2.137E+12</td>
<td>0.0673</td>
<td>STATE</td>
<td>+VELECTRC</td>
</tr>
<tr>
<td>7</td>
<td>74</td>
<td>74</td>
<td>2</td>
<td>2.706E+04</td>
<td>4.068E+13</td>
<td>0.1976</td>
<td>OWNDWEPQ</td>
<td>-FSALARYX</td>
</tr>
<tr>
<td>14T</td>
<td>38</td>
<td>38</td>
<td>2</td>
<td>3.714E+04</td>
<td>4.466E+13</td>
<td>0.3394</td>
<td>-</td>
<td>+DMSXCCPQ</td>
</tr>
<tr>
<td>15T</td>
<td>36</td>
<td>36</td>
<td>2</td>
<td>1.651E+04</td>
<td>3.766E+12</td>
<td>0.8758</td>
<td>-</td>
<td>+ECARTKUC</td>
</tr>
</tbody>
</table>

Number of terminal nodes of final tree: 7
Total number of nodes of final tree: 13
Second best split variable (based on curvature test) at root node is REFGEN

Regression tree:
For categorical variable splits, values not in training data go to the right

Node 1: \( \text{CUTENURE} = "2", "6" \)
Node 2: \( \text{FJSSDEDX} \leq 2720.0000 \)
Node 4: \( \text{FINCBTAX} \leq 114750.50 \)
Node 8: \( \text{INTRDVX-mean} = 3863.4422 \)
Node 4: \( \text{FINCBTAX} > 114750.50 \) or NA
Node 9: \( \text{RETSURV} = "1" \)
Node 18: \( \text{INTRDVX-mean} = 26127.783 \)
Node 9: \( \text{RETSURV} /= "1" \)
Node 19: \( \text{INTRDVX-mean} = 82288.430 \)
Node 2: \( \text{FJSSDEDX} > 2720.0000 \) or NA
Node 5: \( \text{INTRDVX-mean} = 6780.2396 \)
Node 1: \( \text{CUTENURE} /= "2", "6" \)
Node 3: \( \text{RENTEQVX} \leq 4374.0000 \) or NA
Node 6: \( \text{INTRDVX-mean} = 2083.6953 \)
Node 3: \( \text{RENTEQVX} > 4374.0000 \)
Node 7: \( \text{OWNDWEPQ} \leq 5530.5000 \)
Node 14: \( \text{INTRDVX-mean} = 37135.317 \)
Node 7: \( \text{OWNDWEPQ} > 5530.5000 \) or NA

Wei-Yin Loh 96 GUIDE manual
Node 15: INTRDVX-mean = 16508.102

Predictor means below are weighted means of cases with no missing values. Regression coefficients are computed from the complete cases.

WARNING: p-values below not adjusted for split search. For a bootstrap solution see:


Node 1: Intermediate node
A case goes into Node 2 if CUTENURE = "2", "6"
CUTENURE mode = "1"
Coefficients of least squares regression function:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>3167.</td>
<td>2.042</td>
<td>0.4246E-01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STOCKYRX</td>
<td>0.1749E-01</td>
<td>11.42</td>
<td>0.000</td>
<td>0.000</td>
<td>0.3617E+06</td>
<td>0.5450E+07</td>
</tr>
</tbody>
</table>

If regressors have missing values, predicted function value = 4720.7960
Predicted values truncated at 1.00000 & 141304.

----------------------------

Node 2: Intermediate node
A case goes into Node 4 if FJSSDEDX <= 2720.0000
FJSSDEDX mean = 3273.8110

----------------------------

Node 4: Intermediate node
A case goes into Node 8 if FINCBTAX <= 114750.50
FINCBTAX mean = 60208.956

----------------------------

Node 8: Terminal node
Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>2892.</td>
<td>8.287</td>
<td>0.9992E-15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALCBEVCQ</td>
<td>19.81</td>
<td>14.76</td>
<td>0.000</td>
<td>0.000</td>
<td>49.04</td>
<td>4670.</td>
</tr>
</tbody>
</table>

If regressors have missing values, predicted function value = 3863.4422
Predicted values truncated at 1.00000 & 141304.

----------------------------

Node 9: Intermediate node
A case goes into Node 18 if RETSURV = "1"
RETSURV mode = "1"

Wei-Yin Loh 97 GUIDE manual
### 6.2 Piecewise simple polynomial LEAST SQUARES REGRESSION: CE DATA

---

#### Node 18: Terminal node
Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.2096E+05</td>
<td>3.886</td>
<td>0.3316E-03</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FUL0ILCQ</td>
<td>181.5</td>
<td>4.388</td>
<td>0.6841E-04</td>
<td>0.000</td>
<td>28.46</td>
<td>944.0</td>
</tr>
</tbody>
</table>

If regressors have missing values, predicted function value = 26127.783
Predicted values truncated at 1.00000 & 141304.

---

#### Node 19: Terminal node
Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.1215E+06</td>
<td>8.524</td>
<td>0.3657E-03</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROYESTX</td>
<td>-0.5119</td>
<td>-6.012</td>
<td>0.1830E-02</td>
<td>24.00</td>
<td>0.1298E+06</td>
<td>0.2300E+06</td>
</tr>
</tbody>
</table>

If regressors have missing values, predicted function value = 97671.707
Predicted values truncated at 1.00000 & 141304.

---

#### Node 5: Terminal node
Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-953.9</td>
<td>-0.9683</td>
<td>0.3337</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ETOTALC</td>
<td>0.7695</td>
<td>16.20</td>
<td>0.000</td>
<td>0.000</td>
<td>0.1005E+05</td>
<td>0.1615E+06</td>
</tr>
</tbody>
</table>

If regressors have missing values, predicted function value = 6780.2396
Predicted values truncated at 1.00000 & 141304.

---

#### Node 3: Intermediate node
A case goes into Node 6 if RENTEQVX <= 4374.0000 or NA
RENTEQVX mean = 2434.6988

---

#### Node 6: Terminal node
Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1875.</td>
<td>7.721</td>
<td>0.2043E-13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VELECTRC</td>
<td>166.1</td>
<td>10.59</td>
<td>0.4441E-15</td>
<td>0.000</td>
<td>1.258</td>
<td>484.0</td>
</tr>
</tbody>
</table>

If regressors have missing values, predicted function value = 2083.6953
Predicted values truncated at 1.00000 & 141304.

---

#### Node 7: Intermediate node
A case goes into Node 14 if OWNDWEPQ <= 5530.5000
OWNDWEPQ mean = 6589.8800

---

#### Node 14: Terminal node
Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.1346E+05</td>
<td>1.500</td>
<td>0.1422</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DMSXCCPQ</td>
<td>112.6</td>
<td>4.300</td>
<td>0.1244E-03</td>
<td>0.000</td>
<td>210.3</td>
<td>1300.</td>
</tr>
</tbody>
</table>

Wei-Yin Loh 98 GUIDE manual
If regressors have missing values, predicted function value = 37135.317
Predicted values truncated at 1.00000 & 141304.

Node 15: Terminal node

Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>7371.</td>
<td>3.372</td>
<td>0.1874E-02</td>
<td>30.23</td>
<td>415.0</td>
<td></td>
</tr>
<tr>
<td>ECARTKUC</td>
<td>302.2</td>
<td>15.49</td>
<td>0.000</td>
<td>30.23</td>
<td>415.0</td>
<td></td>
</tr>
</tbody>
</table>

If regressors have missing values, predicted function value = 16508.102
Predicted values truncated at 1.00000 & 141304.

Proportion of variance (R-squared) explained by tree model: 0.5504

Observed and fitted values are stored in lin.fit
Regressor names and coefficients are stored in lin.var
LaTeX code for tree is in lin.tex
R code is stored in lin.r

The pruned tree (marked with two asterisks) has 7 terminal nodes and a cross-validation estimate of prediction mean squared error of 6.470E+12.

The tree is shown in Figure 10. Below each terminal node are printed the sample size (in italics), the sample mean of INTRDVX and the signed simple linear predictor, with the sign being that of the slope coefficient. Nodes with mean of the d variable above and below the mean at the root node are colored yellow and purple, respectively.

6.2.3 Plots of data

Figure 11 shows plots of the data and fitted regression lines in the terminal nodes of the tree. The plots are drawn using the R code in Figure 12, which reads the files lin.fit and lin.var. The contents of the latter are below. The first row is a header line. Each subsequent row gives the terminal node number, predictor variable name, intercept and slope of the regression line, and lower and upper truncation limits on the predicted values (the latter defaults are the global minimum and maximum observed values of the dependent variable).

<table>
<thead>
<tr>
<th>node</th>
<th>variable</th>
<th>beta0</th>
<th>beta1</th>
<th>lower</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>ALCBEVCQ</td>
<td>2892.</td>
<td>19.81</td>
<td>1.000</td>
<td>0.1413E+6</td>
</tr>
<tr>
<td>18</td>
<td>FULOILCQ</td>
<td>0.2096E+5</td>
<td>181.5</td>
<td>1.000</td>
<td>0.1413E+6</td>
</tr>
<tr>
<td>19</td>
<td>ROYESTX</td>
<td>0.1215E+6</td>
<td>-0.5119</td>
<td>1.000</td>
<td>0.1413E+6</td>
</tr>
<tr>
<td>5</td>
<td>ETOTALC</td>
<td>-953.9</td>
<td>0.7695</td>
<td>1.000</td>
<td>0.1413E+6</td>
</tr>
<tr>
<td>6</td>
<td>VELECTRC</td>
<td>1875.</td>
<td>166.1</td>
<td>1.000</td>
<td>0.1413E+6</td>
</tr>
</tbody>
</table>
Figure 10: GUIDE v.41.0 0.250-SE piecewise simple linear weighted least-squares regression tree (constant fitted to incomplete cases) for predicting INTRDVX. At each split, an observation goes to the left branch if and only if the condition is satisfied. Symbol ‘≤∗’ stands for ‘≤ or missing’. $S_1 = \{2, 6\}$. Sample size (in italics), weighted mean of INTRDVX, and signed name of regressor variable printed below nodes. Terminal nodes with means above and below value of 5130.6 at root node are colored yellow and purple respectively. Asterisk appended to regressor name indicates its slope is significant at the 0.05 level (unadjusted for multiplicity and model fitting). Second best split variable at root node is REFGEN.
Figure 11: Data and regression lines in terminal nodes of tree in Figure 10. If there are missing values in the regressor, a solid red line marks their mean. If there are no missing values, a dashed red line marks the mean of all points in the node.
```r
z <- read.table("ce2021.txt",header=TRUE)
par(mfrow=c(4,2))
z1 <- read.table("lin.fit",header=TRUE)
z2 <- read.table("lin.var",header=TRUE)
nodes <- unique(sort(z1$node))
y <- z$INTRDVX
for(n in nodes){
  gp <- z1$node == n & z1$train == "y"
  vrow <- z2$node == n
  b0 <- z2$beta0[vrow]
  b1 <- z2$beta1[vrow]
  reg <- z2$variable[vrow]
  k <- which(names(z) %in% reg)
  x <- z[,k]
  plot(y[gp] ~ x[gp],xlab=reg,ylab="INTRDVX",col="blue")
  abline(c(b0,b1))
  nomiss <- z1$node == n & z1$train == "y" & !is.na(x)
  if(sum(nomiss) < sum(gp)){
    miss <- z1$node == n & z1$train == "y" & is.na(x)
    abline(h=mean(y[miss]),col="red",lty=1)
  } else {
    abline(h=mean(y[gp]),col="red",lty=2)
  }
  title(paste("Node",n))
}
```

Figure 12: R code for Figure 11
6.3 Stepwise linear

Besides piecewise constant and best simple polynomial, GUIDE can fit a multiple linear (where all n and f variables are used as regressors) or a stepwise linear (where forward and backward selection is used to select a subset of regressors) regression model at each node. Quite often, these models have higher prediction accuracy, as hinted by the cross-validation estimates of MSE in the output.

For stepwise regression, missing values in each x variable are imputed with the mean of x in the node and a stepwise linear regression model is fitted to the y variable, using the imputed x variables and their missing-value indicators. The name of the indicator of x denoted by x.NA. In other words, x.NA = I(x = NA).

6.3.1 Input file creation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: step.in
   Input 1 for model fitting, 2 for importance or DIF scoring,
   3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: step.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1): 2
   Choose type of regression model:
   1=linear, 2=quantile, 3=Poisson, 4=censored response,
   5=multiresponse or item response, 6=longitudinal data (with T variables),
   7=binary logistic regression.
   Input choice ([1:7], <cr>=1):
   Input 1 for least squares, 2 least median of squares ([1:2], <cr>=1):
   Choose complexity of model to use at each node:
   Choose 0 for stepwise linear regression (recommended for prediction)
   Choose 1 for multiple regression
   Choose 2 for best simple polynomial in one N or F variable
   Choose 3 for constant fit (recommended for interpretability or if there is an R variable)
   0: stepwise linear, 1: multiple linear, 2: best simple polynomial, 3: constant,
   4: best simple stepwise ANCOVA ([0:4], <cr>=3): 0
   Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
   Input name of data description file (max 100 characters);
6.3 Stepwise linear

enclose with matching quotes if it has spaces: ce2021reg.dsc
Reading data description file ...
Training sample file: ce2021.txt
Missing value code: NA
Records in data file start on line 2
Number of M variables associated with C variables: 19
D variable is INTRDVX
Reading data file ...
Number of records in data file: 3965
Length of longest entry in data file: 11
Checking for missing values ...
Finished checking
Missing values found in D variable
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Finding number of levels of M variables associated with C variables ...
Assigning integer codes to values of 47 categorical variables
Finished assigning codes to 10 categorical variables
Finished assigning codes to 20 categorical variables
Finished assigning codes to 30 categorical variables
Finished assigning codes to 40 categorical variables
Associating missing values of N and S variables with M variable codes ...
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...
Data checks complete
Creating missing value indicators ...
Rereading data ...
Warning: N variable DIRACC is constant
Warning: N variable TOTHVHRP is constant
Warning: N variable TOTHVHRC is constant
Warning: N variable ROTHNFLC is constant
Warning: N variable WELFREBX is constant
Warning: N variable OTHLYRBX is constant
Warning: N variable OTHLNYRB is constant
Smallest positive weight: 1.0725E+03
Largest positive weight: 9.3902E+04

<table>
<thead>
<tr>
<th>Total #cases w/ #missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>#cases miss. D ord. vals</td>
</tr>
<tr>
<td>3965 1478 3965 1 384 84 0</td>
</tr>
<tr>
<td>#P-var #M-var #B-var #C-var #I-var</td>
</tr>
<tr>
<td>0 116 0 47 0</td>
</tr>
</tbody>
</table>

Weight variable FINLWT21 in column: 31
Number of cases used for training: 2487
Number of split variables: 431
6.3 Stepwise linear LEAST SQUARES REGRESSION: CE DATA

Number of cases excluded due to 0 W or missing D variable: 1478
Finished reading data file
Input 1 for unweighted, 2 for weighted error estimates during pruning ([1:2], <cr>=2):
Warning: No interaction tests; too many predictor variables
Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):
Input file name to store LaTeX code (use .tex as suffix): step.tex
You can store the variables and/or values used to split and fit in a file
Choose 1 to skip this step, 2 to store split and fit variables,
3 to store split variables and their values
Input your choice ([1:3], <cr>=1):
Input 2 to save regressor names in a file, 1 otherwise ([1:2], <cr>=2):
Input file name: step.reg
Input 2 to save fitted values and node IDs, 1 otherwise ([1:2], <cr>=2):
Input name of file to store node ID and fitted value of each case: step.fit
Input 2 to write R function for predicting new cases, 1 otherwise ([1:2], <cr>=2):
Input file name: step.r
Input rank of top variable to split root node ([1:515], <cr>=1):
Input file is created!
Run GUIDE with the command: guide < step.in

6.3.2 Results

Least squares regression tree
Predictions truncated at global min. and max. of D sample values
Pruning by cross-validation
Data description file: ce2021reg.dsc
Training sample file: ce2021.txt
Missing value code: NA
Records in data file start on line 2
Number of M variables associated with C variables: 19
D variable is INTRDVX
Piecewise forward and backward stepwise regression
F-to-enter and F-to-delete: 4.000 3.990
Using as many variables as needed
Number of records in data file: 3965
Length of longest entry in data file: 11
Missing values found in D variable
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Warning: N variable DIRACC is constant
Warning: N variable TOVHRBP is constant
Warning: N variable TOTHVHRP is constant
Warning: N variable TOTHVHRC is constant
Warning: N variable ROVHRFLC is constant
Warning: N variable WELFREBX is constant
Warning: N variable OTHLYRBX is constant
Warning: N variable OTHLNYRB is constant
Smallest and largest positive weights are 1.0725E+03 and 9.3902E+04

Summary information for training sample of size 2487 (excluding observations with non-positive weight or missing values in d, e, t, r or z variables)
d=dependent, b=split and fit cat variable using indicator variables, c=split-only categorical, i=fit-only categorical (via indicators), s=split-only numerical, n=split and fit numerical, f=fit-only numerical, m=missing-value flag variable, p=periodic variable, w=weight
Levels of M variables are for missing values in associated variables

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIRACC</td>
<td>1.0000E+00</td>
<td>1.000</td>
<td>125</td>
<td></td>
</tr>
<tr>
<td>DIRACC_</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE_REF</td>
<td>19.00</td>
<td>87.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE_REF_</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE2</td>
<td>2.1000E+01</td>
<td>87.00</td>
<td>1092</td>
<td></td>
</tr>
<tr>
<td>AGE2_</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WHOLIFX.NA</td>
<td>0.000</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VEHICTAX.NA</td>
<td>0.000</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CREDYR.NA</td>
<td>0.000</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total #cases w/ #missing
#cases miss. D ord. vals #X-var #N-var #F-var #S-var
3965 1478 3965 1 384 84 0
#P-var #M-var #B-var #C-var #I-var
0 116 0 47 0

Weight variable FINLWT21 in column: 31
Number of cases used for training: 2487
Number of split variables: 431
Number of cases excluded due to 0 W or missing D variable: 1478

Missing regressors imputed with means and missing-value indicators added
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE’s for pruned tree: 0.2500

Weighted error estimates used for pruning
Warning: No interaction tests; too many predictor variables
No nodewise interaction tests
Fraction of cases used for splitting each node: 1.0000
Maximum number of split levels: 12
Minimum node sample size: 20
6.3 Stepwise linear

Top-ranked variables and 1-df chi-squared values at root node
1 0.6230E+02 RETSURV
2 0.5638E+02 AGE_REF
3 0.5312E+02 CUTENURE
4 0.5223E+02 REFGEN

Size and CV MSE and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Nodes</th>
<th>Mean MSE</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median MSE</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>27</td>
<td>5.221E+12</td>
<td>7.323E+11</td>
<td>1.103E+12</td>
<td>3.700E+12</td>
<td>2.079E+12</td>
</tr>
<tr>
<td>6</td>
<td>26</td>
<td>5.225E+12</td>
<td>7.329E+11</td>
<td>1.095E+12</td>
<td>3.915E+12</td>
<td>2.068E+12</td>
</tr>
<tr>
<td>7</td>
<td>24</td>
<td>5.218E+12</td>
<td>7.326E+11</td>
<td>1.098E+12</td>
<td>3.869E+12</td>
<td>2.028E+12</td>
</tr>
<tr>
<td>8</td>
<td>23</td>
<td>5.183E+12</td>
<td>7.321E+11</td>
<td>1.099E+12</td>
<td>3.695E+12</td>
<td>2.025E+12</td>
</tr>
<tr>
<td>9</td>
<td>22</td>
<td>5.191E+12</td>
<td>7.321E+11</td>
<td>1.099E+12</td>
<td>3.703E+12</td>
<td>2.024E+12</td>
</tr>
<tr>
<td>10</td>
<td>21</td>
<td>5.185E+12</td>
<td>7.320E+11</td>
<td>1.099E+12</td>
<td>3.677E+12</td>
<td>2.025E+12</td>
</tr>
<tr>
<td>11</td>
<td>17</td>
<td>5.073E+12</td>
<td>7.213E+11</td>
<td>1.123E+12</td>
<td>3.413E+12</td>
<td>2.120E+12</td>
</tr>
<tr>
<td>12</td>
<td>15</td>
<td>5.074E+12</td>
<td>7.213E+11</td>
<td>1.123E+12</td>
<td>3.413E+12</td>
<td>2.119E+12</td>
</tr>
<tr>
<td>13</td>
<td>13</td>
<td>5.078E+12</td>
<td>7.212E+11</td>
<td>1.121E+12</td>
<td>3.413E+12</td>
<td>2.119E+12</td>
</tr>
<tr>
<td>14+</td>
<td>11</td>
<td>4.974E+12</td>
<td>7.145E+11</td>
<td>1.132E+12</td>
<td>2.891E+12</td>
<td>2.156E+12</td>
</tr>
<tr>
<td>15</td>
<td>10</td>
<td>5.051E+12</td>
<td>7.084E+11</td>
<td>1.170E+12</td>
<td>2.891E+12</td>
<td>2.217E+12</td>
</tr>
<tr>
<td>16++</td>
<td>9</td>
<td>4.892E+12</td>
<td>7.016E+11</td>
<td>1.129E+12</td>
<td>3.068E+12</td>
<td>2.162E+12</td>
</tr>
<tr>
<td>17</td>
<td>8</td>
<td>5.043E+12</td>
<td>6.730E+11</td>
<td>1.029E+12</td>
<td>4.422E+12</td>
<td>1.489E+12</td>
</tr>
<tr>
<td>18</td>
<td>6</td>
<td>5.286E+12</td>
<td>7.237E+11</td>
<td>1.134E+12</td>
<td>4.797E+12</td>
<td>2.107E+12</td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 28 terminal nodes
0-SE tree based on median is marked with + and has 11 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on median and bootstrap SE is marked with ++
** tree same as -- tree
* tree same as ** tree
* tree same as -- tree

Following tree is based on mean CV with naive SE estimate (**)

Structure of final tree. Each terminal node is marked with a T.

D-mean is weighted mean of INTRDVX in the node

Wei-Yin Loh
GUIDE manual
### 6.3 Stepwise Linear Least Squares Regression: CE Data

Cases fit give the number of cases used to fit node
MSE and $R^2$ are based on all cases in node

<table>
<thead>
<tr>
<th>Node</th>
<th>Total cases</th>
<th>Cases fit</th>
<th>Rank</th>
<th>D-mean</th>
<th>MSE</th>
<th>$R^2$</th>
<th>Split variable</th>
<th>Other variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2487</td>
<td>2487</td>
<td>26</td>
<td>5.131E+03</td>
<td>7.866E+12</td>
<td>0.0603</td>
<td>RETSURV</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>651</td>
<td>651</td>
<td>24</td>
<td>5.786E+03</td>
<td>6.598E+12</td>
<td>0.0804</td>
<td>OTHREGX</td>
<td></td>
</tr>
<tr>
<td>4T</td>
<td>28</td>
<td>28</td>
<td>4</td>
<td>2.938E+03</td>
<td>5.458E+10</td>
<td>0.9715</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>623</td>
<td>623</td>
<td>24</td>
<td>5.926E+03</td>
<td>6.754E+12</td>
<td>0.0902</td>
<td>SEX2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>244</td>
<td>244</td>
<td>20</td>
<td>1.033E+04</td>
<td>4.946E+12</td>
<td>0.6703</td>
<td>FINDRETX</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>217</td>
<td>217</td>
<td>15</td>
<td>9.797E+03</td>
<td>5.371E+12</td>
<td>0.5974</td>
<td>EOTHLODC</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>193</td>
<td>193</td>
<td>15</td>
<td>9.425E+03</td>
<td>5.011E+12</td>
<td>0.6497</td>
<td>ALLFULPQ</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>173</td>
<td>173</td>
<td>11</td>
<td>9.107E+03</td>
<td>5.570E+12</td>
<td>0.5629</td>
<td>ETRANTP</td>
<td></td>
</tr>
<tr>
<td>160</td>
<td>46</td>
<td>46</td>
<td>8</td>
<td>3.679E+03</td>
<td>8.093E+12</td>
<td>0.5962</td>
<td>ERANKHM</td>
<td></td>
</tr>
<tr>
<td>320T</td>
<td>21</td>
<td>21</td>
<td>2</td>
<td>6.075E+03</td>
<td>1.276E+12</td>
<td>0.6159</td>
<td></td>
<td></td>
</tr>
<tr>
<td>321T</td>
<td>25</td>
<td>25</td>
<td>1</td>
<td>1.713E+03</td>
<td>1.0000E+00</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>322</td>
<td>71</td>
<td>71</td>
<td>5</td>
<td>1.607E+04</td>
<td>9.315E+12</td>
<td>0.5794</td>
<td>RETSURVX</td>
<td></td>
</tr>
<tr>
<td>644T</td>
<td>44</td>
<td>44</td>
<td>9</td>
<td>1.334E+04</td>
<td>2.211E+12</td>
<td>0.8700</td>
<td>HEALTHCQ</td>
<td></td>
</tr>
<tr>
<td>645T</td>
<td>27</td>
<td>27</td>
<td>5</td>
<td>2.023E+04</td>
<td>1.200E+13</td>
<td>0.6888</td>
<td></td>
<td></td>
</tr>
<tr>
<td>323</td>
<td>56</td>
<td>56</td>
<td>9</td>
<td>4.610E+03</td>
<td>8.122E+11</td>
<td>0.9030</td>
<td>MRTINTPQ</td>
<td></td>
</tr>
<tr>
<td>646T</td>
<td>36</td>
<td>36</td>
<td>4</td>
<td>3.565E+03</td>
<td>4.629E+11</td>
<td>0.6672</td>
<td></td>
<td></td>
</tr>
<tr>
<td>647T</td>
<td>20</td>
<td>20</td>
<td>5</td>
<td>6.682E+03</td>
<td>1.754E+09</td>
<td>0.9999</td>
<td></td>
<td></td>
</tr>
<tr>
<td>81T</td>
<td>20</td>
<td>20</td>
<td>2</td>
<td>1.286E+04</td>
<td>2.110E+13</td>
<td>0.2045</td>
<td></td>
<td></td>
</tr>
<tr>
<td>41T</td>
<td>24</td>
<td>24</td>
<td>4</td>
<td>1.285E+04</td>
<td>2.842E+12</td>
<td>0.6027</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21T</td>
<td>27</td>
<td>27</td>
<td>13</td>
<td>1.449E+04</td>
<td>2.975E+11</td>
<td>0.9936</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11T</td>
<td>379</td>
<td>379</td>
<td>41</td>
<td>3.114E+03</td>
<td>4.841E+11</td>
<td>0.8217</td>
<td>UNISTRQ</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1836</td>
<td>1836</td>
<td>22</td>
<td>4.915E+03</td>
<td>7.982E+12</td>
<td>0.1003</td>
<td>OTHRINCX</td>
<td></td>
</tr>
<tr>
<td>6T</td>
<td>36</td>
<td>36</td>
<td>3</td>
<td>3.484E+02</td>
<td>2.533E+10</td>
<td>0.1177</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1800</td>
<td>1800</td>
<td>22</td>
<td>4.971E+03</td>
<td>8.307E+12</td>
<td>0.0817</td>
<td>STATE</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>160</td>
<td>160</td>
<td>16</td>
<td>2.234E+03</td>
<td>1.719E+11</td>
<td>0.7542</td>
<td>EDUCAQP</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>140</td>
<td>140</td>
<td>10</td>
<td>2.335E+03</td>
<td>2.460E+11</td>
<td>0.6703</td>
<td>EMRTINP</td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>104</td>
<td>104</td>
<td>13</td>
<td>1.762E+03</td>
<td>8.363E+10</td>
<td>0.8491</td>
<td>OCCUCOD1</td>
<td></td>
</tr>
<tr>
<td>112T</td>
<td>36</td>
<td>36</td>
<td>10</td>
<td>3.930E+03</td>
<td>2.080E+11</td>
<td>0.8734</td>
<td></td>
<td></td>
</tr>
<tr>
<td>113</td>
<td>68</td>
<td>68</td>
<td>6</td>
<td>5.372E+02</td>
<td>5.714E+09</td>
<td>0.5147</td>
<td>FOODPQ</td>
<td></td>
</tr>
<tr>
<td>226T</td>
<td>39</td>
<td>39</td>
<td>13</td>
<td>6.323E+02</td>
<td>4.990E+08</td>
<td>0.9750</td>
<td></td>
<td></td>
</tr>
<tr>
<td>227T</td>
<td>29</td>
<td>29</td>
<td>4</td>
<td>4.123E+02</td>
<td>4.663E+09</td>
<td>0.3912</td>
<td></td>
<td></td>
</tr>
<tr>
<td>57T</td>
<td>36</td>
<td>36</td>
<td>10</td>
<td>4.059E+03</td>
<td>1.084E+11</td>
<td>0.9346</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29T</td>
<td>20</td>
<td>20</td>
<td>2</td>
<td>1.580E+03</td>
<td>1.243E+11</td>
<td>0.3503</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1640</td>
<td>1640</td>
<td>22</td>
<td>5.255E+03</td>
<td>9.015E+12</td>
<td>0.0854</td>
<td>INCDNW1</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>291</td>
<td>291</td>
<td>31</td>
<td>1.590E+04</td>
<td>8.301E+12</td>
<td>0.7299</td>
<td>FRRERTRM</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>191</td>
<td>191</td>
<td>23</td>
<td>1.370E+04</td>
<td>5.513E+12</td>
<td>0.7716</td>
<td>BUILT</td>
<td></td>
</tr>
<tr>
<td>120</td>
<td>131</td>
<td>131</td>
<td>13</td>
<td>1.784E+04</td>
<td>4.897E+12</td>
<td>0.8436</td>
<td>PERSACQ</td>
<td></td>
</tr>
<tr>
<td>240</td>
<td>109</td>
<td>109</td>
<td>12</td>
<td>1.671E+04</td>
<td>2.554E+12</td>
<td>0.9080</td>
<td>ERANKH</td>
<td></td>
</tr>
<tr>
<td>480</td>
<td>81</td>
<td>81</td>
<td>12</td>
<td>1.153E+04</td>
<td>1.421E+12</td>
<td>0.9345</td>
<td>FOODPQ</td>
<td></td>
</tr>
<tr>
<td>960T</td>
<td>43</td>
<td>43</td>
<td>10</td>
<td>8.826E+03</td>
<td>4.548E+10</td>
<td>0.9979</td>
<td>STATE</td>
<td></td>
</tr>
<tr>
<td>961T</td>
<td>38</td>
<td>38</td>
<td>7</td>
<td>1.541E+04</td>
<td>1.748E+12</td>
<td>0.9288</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
6.3 Stepwise linear 6 LEAST SQUARES REGRESSION: CE DATA

481T 28 28 7 3.204E+04 5.745E+12 0.8795 -
241T 22 22 4 2.354E+04 1.333E+13 0.7562 -
121 60 60 11 5.283E+03 3.681E+11 0.9132 NUM_AUTO
242T 36 36 5 2.372E+03 2.014E+11 0.7755 -
243T 24 24 8 9.103E+03 1.956E+10 0.9981 -
61 100 100 8 2.028E+04 1.873E+12 0.9555 BUILT
122T 37 37 7 1.479E+04 8.768E+10 0.9977 -
123 63 63 7 2.343E+04 2.312E+12 0.9520 HEALTHPQ
246T 27 27 5 7.647E+03 9.147E+10 0.9933 -
247 36 36 8 3.517E+04 2.369E+12 0.9690 -
31 1349 1349 20 3.085E+03 4.887E+12 0.0800 FSALARYX
62T 434 434 25 3.001E+03 4.335E+09 0.9993 -
63T 915 915 18 3.124E+03 1.804E+08 1.0000 -

Warning: tree very large, omitting node numbers in LaTeX file

Number of terminal nodes of final tree: 28
Total number of nodes of final tree: 55
Second best split variable (based on curvature test) at root node is AGE_REF

Regression tree:
For categorical variable splits, values not in training data go to the right

Node 1: RETSURV = "1"
  Node 2: OTHREGX <= 11337.000
    Node 4: INTRDVX-mean = 2938.2174
  Node 2: OTHREGX > 11337.000 or NA
    Node 5: SEX2 = "2"
      Node 10: FINDRETX <= 2350.0000
        Node 20: EOTHLODC <= 575.00000
          Node 40: ALLFULPQ <= 41.00000
            Node 80: ETRANPTP <= 444.16650
              Node 160: ERANKHM <= 0.51617950
                Node 320: INTRDVX-mean = 6075.2030
                  Node 160: ERANKHM > 0.51617950 or NA
                    Node 321: INTRDVX-mean = 1712.6607
                      Node 80: ETRANPTP > 444.16650 or NA
                        Node 161: CUTENURE = "2"
                          Node 322: RETSURVX <= 31600.000
                            Node 644: INTRDVX-mean = 13336.430
                              Node 642: RETSURVX > 31600.000 or NA
                                Node 645: INTRDVX-mean = 20229.476
                                  Node 161: CUTENURE /= "2"
                                    Node 323: MRTINTPQ <= 970.00000
                                      Node 646: INTRDVX-mean = 3564.6625
                                        Node 323: MRTINTPQ > 970.00000 or NA
                                          Node 647: INTRDVX-mean = 6681.5390

Wei-Yin Loh 109 GUIDE manual
6.3  Stepwise linear  LEAST SQUARES REGRESSION: CE DATA

Node 40: ALFLUPQ > 41.000000 or NA
  Node 81: INTRDVX-mean = 12860.244
  Node 20: EOTHLODC > 575.000000 or NA
    Node 41: INTRDVX-mean = 12851.752
  Node 10: FINDRETX > 2350.0000 or NA
    Node 21: INTRDVX-mean = 14488.623
  Node 5: SEX2 /= "2"
    Node 11: INTRDVX-mean = 3114.4256
Node 1: RETSURV /= "1"
  Node 3: OTHRINCX <= 4083.0000
    Node 6: INTRDVX-mean = 348.37734
  Node 3: OTHRINCX > 4083.0000 or NA
    Node 7: STATE = "10", "22", "40", "41", "53"
    Node 14: EDUCAPQ <= 162.00000
      Node 28: EMRTPNOP <= 991.50000
        Node 56: OCCUCOD1 = "1"
          or (OCCUCOD1 = NA & OCCU_OD1 = "A")
        Node 112: INTRDVX-mean = 3929.7658
      Node 56: OCCUCOD1 /= "1"
        & not (OCCUCOD1 = NA & OCCU_OD1 = "A")
      Node 113: FOODPQ <= 1282.6667
      Node 226: INTRDVX-mean = 632.29761
      Node 113: FOODPQ > 1282.6667 or NA
      Node 227: INTRDVX-mean = 412.30881
    Node 28: EMRTPNOP > 991.50000 or NA
    Node 57: INTRDVX-mean = 4059.2022
  Node 14: EDUCAPQ > 162.000000 or NA
    Node 29: INTRDVX-mean = 1580.4405
    Node 7: STATE /= "10", "22", "40", "41", "53"
    Node 15: INCNONW1 = "1"
    Node 30: FRRETIRM <= 29823.000
      Node 60: BUILT <= 1999.5000
        Node 120: PERSCACQ <= 50.000000
          Node 240: ERANKH <= 0.55780640
            Node 480: FOODPQ <= 686.66665
              Node 960: INTRDVX-mean = 8825.6885
              Node 480: FOODPQ > 686.66665 or NA
              Node 961: INTRDVX-mean = 15411.105
            Node 240: ERANKH > 0.55780640 or NA
              Node 481: INTRDVX-mean = 32038.502
            Node 120: PERSCACQ > 50.000000 or NA
            Node 241: INTRDVX-mean = 23538.397
        Node 120: BUILT > 1999.5000 or NA
          Node 121: NUM_AUTO <= 0.50000000
            Node 242: INTRDVX-mean = 2372.2425
            Node 121: NUM_AUTO > 0.50000000 or NA
6.3 Stepwise linear LEAST SQUARES REGRESSION: CE DATA

Node 243: INTRDVX-mean = 9103.208

Node 30: FRRETIRM > 29823.000 or NA

Node 61: BUILT <= 1962.500

Node 122: INTRDVX-mean = 14794.456

Node 61: BUILT > 1962.500 or NA

Node 123: HEALTHPQ <= 1336.000

Node 246: INTRDVX-mean = 7646.6109

Node 123: HEALTHPQ > 1336.000 or NA

Node 247: INTRDVX-mean = 35168.090

Node 15: INCNONW1 /= "1"

Node 31: FSALARYX <= 60200.000

Node 62: INTRDVX-mean = 3000.6016

Node 31: FSALARYX > 60200.000 or NA

Node 63: INTRDVX-mean = 3123.6922

**********************************************************************
Predictor means below are weighted means of cases with no missing values.
Regression coefficients are computed from the complete cases.

WARNING: p-values below not adjusted for split search. For a bootstrap solution see:

1. Loh et al. (2016), "Identification of subgroups with differential treatment effects
   for longitudinal and multiresponse variables", Statistics in Medicine, v.35, 4837-4855.

2. Loh et al. (2019), "Subgroups from regression trees with adjustment for prognostic
   effects and post-selection inference", Statistics in Medicine, v.38, 545-557.

   in "Design and Analysis of Subgroups with Biopharmaceutical Applications", Springer, pp.147-165.

Node 1: Intermediate node
A case goes into Node 2 if RETSURV = "1"
RETSURV mode = "2"

Coefficients of least squares regression function:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.3652E+05</td>
<td>-0.8428</td>
<td>0.3994</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FRRETIRX</td>
<td>0.2151</td>
<td>7.734</td>
<td>0.000</td>
<td>0.000</td>
<td>0.1071E+05</td>
<td>0.1116E+06</td>
</tr>
<tr>
<td>FSALARYX</td>
<td>0.7259E-02</td>
<td>1.817</td>
<td>0.6927E-01</td>
<td>0.000</td>
<td>0.8172E+05</td>
<td>0.7645E+06</td>
</tr>
<tr>
<td>FSSIX</td>
<td>0.4632E-01</td>
<td>0.1486</td>
<td>0.8818</td>
<td>0.000</td>
<td>83.02</td>
<td>0.3600E+05</td>
</tr>
<tr>
<td>OTHRINRX</td>
<td>0.3675E-01</td>
<td>0.5329</td>
<td>0.5941</td>
<td>150.0</td>
<td>0.2064E+05</td>
<td>0.1041E+06</td>
</tr>
<tr>
<td>FSMPFRMIX</td>
<td>0.4211E-01</td>
<td>4.769</td>
<td>0.1956E-05</td>
<td>-0.1160E+06</td>
<td>6641.</td>
<td>0.7703E+06</td>
</tr>
<tr>
<td>JFS_AMT</td>
<td>0.7798E-01</td>
<td>1.264</td>
<td>0.8994</td>
<td>0.000</td>
<td>64.43</td>
<td>9600.</td>
</tr>
<tr>
<td>NETRENTX</td>
<td>-0.1567E-01</td>
<td>-0.3371</td>
<td>0.7361</td>
<td>-0.1402E+05</td>
<td>0.1462E+05</td>
<td>0.1589E+06</td>
</tr>
<tr>
<td>NETRNTBX</td>
<td>-0.1043</td>
<td>-0.3341</td>
<td>0.7383</td>
<td>-2400.</td>
<td>9325.</td>
<td>0.7130E+05</td>
</tr>
<tr>
<td>OTHRERBGX</td>
<td>0.1112</td>
<td>0.2365</td>
<td>0.8131</td>
<td>488.0</td>
<td>6380.</td>
<td>0.4200E+05</td>
</tr>
<tr>
<td>OTHRREGX</td>
<td>0.1041</td>
<td>1.332</td>
<td>0.1829</td>
<td>100.0</td>
<td>0.1281E+05</td>
<td>0.8288E+05</td>
</tr>
</tbody>
</table>

Wei-Yin Loh 111 GUIDE manual
### 6.3 Stepwise linear LEAST SQUARES REGRESSION: CE DATA

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.6644E+05</td>
<td>19.05</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FINCBTAX</td>
<td>0.9992</td>
<td>623.2</td>
<td>0.000</td>
<td>-0.1050E+06</td>
<td>0.5582E+05</td>
<td>0.5265E+06</td>
</tr>
<tr>
<td>FARETIRX</td>
<td>-0.9981</td>
<td>-373.3</td>
<td>0.000</td>
<td>0.000</td>
<td>5260.</td>
<td>0.9000E+05</td>
</tr>
<tr>
<td>FSALARYX</td>
<td>-0.9983</td>
<td>-526.6</td>
<td>0.000</td>
<td>0.000</td>
<td>0.2803E+05</td>
<td>0.6000E+05</td>
</tr>
<tr>
<td>FSSIX</td>
<td>-0.8487</td>
<td>-59.82</td>
<td>0.000</td>
<td>0.000</td>
<td>171.0</td>
<td>0.3600E+05</td>
</tr>
<tr>
<td>OTHRINCX</td>
<td>-0.9928</td>
<td>-44.32</td>
<td>0.000</td>
<td>4166.</td>
<td>0.1168E+05</td>
<td>0.2000E+05</td>
</tr>
<tr>
<td>ENOMOTRC</td>
<td>-64.69</td>
<td>-24.33</td>
<td>0.000</td>
<td>0.000</td>
<td>1.561</td>
<td>372.0</td>
</tr>
<tr>
<td>FSMPFRMX</td>
<td>-0.9991</td>
<td>-535.9</td>
<td>0.000</td>
<td>-0.1160E+06</td>
<td>0.1448E+05</td>
<td>0.3852E+06</td>
</tr>
<tr>
<td>IRABX</td>
<td>-0.1430</td>
<td>-39.57</td>
<td>0.000</td>
<td>5000.</td>
<td>0.7002E+05</td>
<td>0.3250E+06</td>
</tr>
<tr>
<td>IRAYRB</td>
<td>-0.2897E+05</td>
<td>-24.59</td>
<td>0.000</td>
<td>2.000</td>
<td>3.377</td>
<td>5.000</td>
</tr>
<tr>
<td>IRAYRBX</td>
<td>0.3323</td>
<td>32.66</td>
<td>0.000</td>
<td>5000.</td>
<td>0.8364E+05</td>
<td>0.3250E+06</td>
</tr>
<tr>
<td>JFS_AMT</td>
<td>-1.012</td>
<td>-62.19</td>
<td>0.000</td>
<td>0.000</td>
<td>285.2</td>
<td>9600.</td>
</tr>
<tr>
<td>LIQUIDBX</td>
<td>-0.1103</td>
<td>-11.91</td>
<td>0.000</td>
<td>100.0</td>
<td>0.1477E+05</td>
<td>0.7500E+05</td>
</tr>
<tr>
<td>LIQUYRDB</td>
<td>517.6</td>
<td>4.919</td>
<td>0.000</td>
<td>0.1261E-05</td>
<td>1.000</td>
<td>3.556</td>
</tr>
<tr>
<td>NETRENTX</td>
<td>-0.9983</td>
<td>-292.9</td>
<td>0.000</td>
<td>-0.1402E+05</td>
<td>0.1433E+05</td>
<td>0.1047E+06</td>
</tr>
<tr>
<td>OTHREGB</td>
<td>-0.9985</td>
<td>-6.880</td>
<td>0.000</td>
<td>488.0</td>
<td>1802.</td>
<td>6420.</td>
</tr>
<tr>
<td>OTHREG</td>
<td>-0.9996</td>
<td>-248.5</td>
<td>0.000</td>
<td>144.0</td>
<td>0.1341E+05</td>
<td>0.8288E+05</td>
</tr>
<tr>
<td>ROYESTX</td>
<td>-0.9992</td>
<td>-367.6</td>
<td>0.000</td>
<td>5.000</td>
<td>0.2532E+05</td>
<td>0.2300E+06</td>
</tr>
<tr>
<td>OTHRINCX NA</td>
<td>0.1166E+05</td>
<td>84.11</td>
<td>0.000</td>
<td>0.000</td>
<td>0.9745</td>
<td>1.000</td>
</tr>
</tbody>
</table>

---

Node 2: Intermediate node
A case goes into Node 4 if OTHREGX <= 11337.000
OTHREGX mean = 12453.689

Node 62: Terminal node

Coefficients of least squares regression functions:

---

Wei-Yin Loh

112 GUIDE manual
6.3 Stepwise linear

LEAST SQUARES REGRESSION: CE DATA

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-6543.</td>
<td>-22.65</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FINCBTAX</td>
<td>0.9981</td>
<td>4925.</td>
<td>0.00</td>
<td>0.2884E+05</td>
<td>0.1760E+06</td>
<td>0.1039E+07</td>
</tr>
<tr>
<td>FRRETIRX</td>
<td>-0.9982</td>
<td>-1909.</td>
<td>0.00</td>
<td>0.00</td>
<td>1016.</td>
<td>0.6000E+05</td>
</tr>
<tr>
<td>FSALARYX</td>
<td>-0.9981</td>
<td>-4764.</td>
<td>0.00</td>
<td>0.6040E+05</td>
<td>0.1624E+06</td>
<td>0.7645E+06</td>
</tr>
<tr>
<td>FSSIX</td>
<td>-1.009</td>
<td>-318.8</td>
<td>0.00</td>
<td>0.00</td>
<td>43.16</td>
<td>0.2400E+05</td>
</tr>
<tr>
<td>OTHRINCX</td>
<td>-0.9968</td>
<td>-605.5</td>
<td>0.00</td>
<td>5000.</td>
<td>0.1821E+05</td>
<td>0.5640E+05</td>
</tr>
<tr>
<td>FSPFIRNX</td>
<td>-0.9981</td>
<td>-4690.</td>
<td>0.00</td>
<td>-0.1160E+06</td>
<td>6606.</td>
<td>0.7703E+06</td>
</tr>
<tr>
<td>NETRENTX</td>
<td>-0.9978</td>
<td>-2543.</td>
<td>0.00</td>
<td>-0.1402E+05</td>
<td>0.1458E+05</td>
<td>0.1589E+06</td>
</tr>
<tr>
<td>OTHREG</td>
<td>-1965.</td>
<td>-174.2</td>
<td>0.00</td>
<td>1.00</td>
<td>3.121</td>
<td>7.00</td>
</tr>
<tr>
<td>OTHREGX</td>
<td>-0.9989</td>
<td>-832.9</td>
<td>0.00</td>
<td>400.0</td>
<td>9303.</td>
<td>0.4000E+05</td>
</tr>
<tr>
<td>ROYESTX</td>
<td>-0.9984</td>
<td>-1761.</td>
<td>0.00</td>
<td>10.00</td>
<td>0.2781E+05</td>
<td>0.8700E+05</td>
</tr>
<tr>
<td>OTHRINCX.NA</td>
<td>0.1823E+05</td>
<td>598.6</td>
<td>0.00</td>
<td>0.00</td>
<td>0.9908</td>
<td>1.00</td>
</tr>
<tr>
<td>WELFREBX.NA</td>
<td>0.1069E+05</td>
<td>39.55</td>
<td>0.00</td>
<td>0.00</td>
<td>0.9999</td>
<td>1.00</td>
</tr>
<tr>
<td>NETRENTB.NA</td>
<td>-2389.</td>
<td>-34.05</td>
<td>0.00</td>
<td>0.00</td>
<td>0.9983</td>
<td>1.00</td>
</tr>
<tr>
<td>NETRENTX.NA</td>
<td>0.1458E+05</td>
<td>1271.</td>
<td>0.00</td>
<td>0.00</td>
<td>0.9202</td>
<td>1.00</td>
</tr>
<tr>
<td>OTHREGSX.NA</td>
<td>4255.</td>
<td>144.9</td>
<td>0.00</td>
<td>0.00</td>
<td>0.9904</td>
<td>1.00</td>
</tr>
<tr>
<td>OTHREGX.NA</td>
<td>9329.</td>
<td>839.8</td>
<td>0.00</td>
<td>0.00</td>
<td>0.9203</td>
<td>1.00</td>
</tr>
<tr>
<td>ROYESTX.NA</td>
<td>0.2775E+05</td>
<td>1394.</td>
<td>0.00</td>
<td>0.00</td>
<td>0.9744</td>
<td>1.00</td>
</tr>
<tr>
<td>INTRDVX mean =</td>
<td>3123.69</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Predicted values truncated at 1.00000 & 141304.

------------------------------------------------

Node 63: Terminal node

Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-6543.</td>
<td>-22.65</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FINCBTAX</td>
<td>0.9981</td>
<td>4925.</td>
<td>0.00</td>
<td>0.2884E+05</td>
<td>0.1760E+06</td>
<td>0.1039E+07</td>
</tr>
<tr>
<td>FRRETIRX</td>
<td>-0.9982</td>
<td>-1909.</td>
<td>0.00</td>
<td>0.00</td>
<td>1016.</td>
<td>0.6000E+05</td>
</tr>
<tr>
<td>FSALARYX</td>
<td>-0.9981</td>
<td>-4764.</td>
<td>0.00</td>
<td>0.6040E+05</td>
<td>0.1624E+06</td>
<td>0.7645E+06</td>
</tr>
<tr>
<td>FSSIX</td>
<td>-1.009</td>
<td>-318.8</td>
<td>0.00</td>
<td>0.00</td>
<td>43.16</td>
<td>0.2400E+05</td>
</tr>
<tr>
<td>OTHRINCX</td>
<td>-0.9968</td>
<td>-605.5</td>
<td>0.00</td>
<td>5000.</td>
<td>0.1821E+05</td>
<td>0.5640E+05</td>
</tr>
<tr>
<td>FSPFIRNX</td>
<td>-0.9981</td>
<td>-4690.</td>
<td>0.00</td>
<td>-0.1160E+06</td>
<td>6606.</td>
<td>0.7703E+06</td>
</tr>
<tr>
<td>NETRENTX</td>
<td>-0.9978</td>
<td>-2543.</td>
<td>0.00</td>
<td>-0.1402E+05</td>
<td>0.1458E+05</td>
<td>0.1589E+06</td>
</tr>
<tr>
<td>OTHREG</td>
<td>-1965.</td>
<td>-174.2</td>
<td>0.00</td>
<td>1.00</td>
<td>3.121</td>
<td>7.00</td>
</tr>
<tr>
<td>OTHREGX</td>
<td>-0.9989</td>
<td>-832.9</td>
<td>0.00</td>
<td>400.0</td>
<td>9303.</td>
<td>0.4000E+05</td>
</tr>
<tr>
<td>ROYESTX</td>
<td>-0.9984</td>
<td>-1761.</td>
<td>0.00</td>
<td>10.00</td>
<td>0.2781E+05</td>
<td>0.8700E+05</td>
</tr>
<tr>
<td>OTHRINCX.NA</td>
<td>0.1823E+05</td>
<td>598.6</td>
<td>0.00</td>
<td>0.00</td>
<td>0.9908</td>
<td>1.00</td>
</tr>
<tr>
<td>WELFREBX.NA</td>
<td>0.1069E+05</td>
<td>39.55</td>
<td>0.00</td>
<td>0.00</td>
<td>0.9999</td>
<td>1.00</td>
</tr>
<tr>
<td>NETRENTB.NA</td>
<td>-2389.</td>
<td>-34.05</td>
<td>0.00</td>
<td>0.00</td>
<td>0.9983</td>
<td>1.00</td>
</tr>
<tr>
<td>NETRENTX.NA</td>
<td>0.1458E+05</td>
<td>1271.</td>
<td>0.00</td>
<td>0.00</td>
<td>0.9202</td>
<td>1.00</td>
</tr>
<tr>
<td>OTHREGSX.NA</td>
<td>4255.</td>
<td>144.9</td>
<td>0.00</td>
<td>0.00</td>
<td>0.9904</td>
<td>1.00</td>
</tr>
<tr>
<td>OTHREGX.NA</td>
<td>9329.</td>
<td>839.8</td>
<td>0.00</td>
<td>0.00</td>
<td>0.9203</td>
<td>1.00</td>
</tr>
<tr>
<td>ROYESTX.NA</td>
<td>0.2775E+05</td>
<td>1394.</td>
<td>0.00</td>
<td>0.00</td>
<td>0.9744</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Predicted values truncated at 1.00000 & 141304.

------------------------------------------------

Proportion of variance (R-squared) explained by tree model: 0.9375

The tree is shown in Figure 13. The contents of step.reg below show for each terminal node, the node number, lower and upper truncation values, and the variables selected by stepwise regression in each node.

node lower upper variables

Wei-Yin Loh

GUIDE manual
7 Quantile regression: CE data

GUIDE can build piecewise-constant and piecewise-linear quantile regression models. First we show how to build a piecewise-constant 0.90-quantile regression model.

7.1 Piecewise constant: one quantile

7.1.1 Input file creation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: quantcon.in
   Input 1 for model fitting, 2 for importance or DIF scoring,
   3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: quantcon.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1): 2
   Choose type of regression model:
   1=linear, 2=quantile, 3=Poisson, 4=censored response,
   5=multiresponse or itemresponse, 6=longitudinal data (with T variables),
   7=binary logistic regression.
   Input choice ([1:7], <cr>=1): 2
   Choose complexity of model to use at each node:
   Choose 1 for multiple regression (recommended for prediction)
   Choose 2 for best simple polynomial in one N or F variable
   Choose 3 for constant fit (recommended for interpretability or if there is an R variable)
   1: multiple linear, 2: best simple polynomial, 3: constant ([1:3], <cr>=3):
   Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
   Input 1 for 1 quantile, 2 for 2 quantiles ([1:2], <cr>=1):
   Input quantile probability ([0.00:1.00], <cr>=0.50): 0.90
   Input name of data description file (max 100 characters);
Figure 13: GUIDE v.41.1 0.250-SE piecewise-stepwise linear weighted least-squares regression tree (missing regressor values imputed and missing indicators added) for predicting INTRDVX. At each split, an observation goes to the left branch if and only if the condition is satisfied. $S_1 = \{10, 22, 40, 41, 53\}$. Sample size (in italics) and weighted mean of INTRDVX printed below nodes. Terminal nodes with means above and below value of 5130.6 at root node are colored yellow and purple respectively. Second best split variable at root node is AGE_REF.
enclose with matching quotes if it has spaces: ce2021reg.dsc
Reading data description file ...
Training sample file: ce2021.txt
Missing value code: NA
Records in data file start on line 2
Number of M variables associated with C variables: 19
384 N variables changed to S
D variable is INTRDVX
Reading data file ...
Number of records in data file: 3965
Length of longest entry in data file: 11
Checking for missing values ...
Finished checking
Missing values found in D variable
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Finding number of levels of M variables associated with C variables ...
Assigning integer codes to values of 47 categorical variables
Finished assigning codes to 10 categorical variables
Finished assigning codes to 20 categorical variables
Finished assigning codes to 30 categorical variables
Finished assigning codes to 40 categorical variables
Associating missing values of N and S variables with M variable codes ...
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...
Data checks complete
Creating missing value indicators ...
Rereading data ...
Warning: S variable DIRACC is constant
Warning: S variable TOTHVHRP is constant
Warning: S variable TOTHVHRC is constant
Warning: S variable ROTHRFLC is constant
Warning: S variable WELFREBX is constant
Warning: S variable OTHLYRBX is constant
Warning: S variable OTHLNYRB is constant
Smallest positive weight: 1.0725E+03
Largest positive weight: 9.3902E+04

<table>
<thead>
<tr>
<th>#cases w/ #missing</th>
<th>#cases miss. D ord. vals</th>
<th>#X-var</th>
<th>#N-var</th>
<th>#F-var</th>
<th>#S-var</th>
<th>#P-var</th>
<th>#M-var</th>
<th>#B-var</th>
<th>#C-var</th>
<th>#I-var</th>
</tr>
</thead>
<tbody>
<tr>
<td>3965</td>
<td>1478</td>
<td>3965</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>384</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>116</td>
<td>0</td>
<td>47</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Number of cases used for training: 2487
Number of split variables: 431
7.1 Piecewise constant: one quantile

Number of cases excluded due to 0 W or missing D variable: 1478
Finished reading data file
Input 1 for unweighted, 2 for weighted error estimates during pruning ([1:2], <cr>=2):
Warning: No interaction tests; too many predictor variables
Warning: All positive weights treated as 1
Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):
Input file name to store LaTeX code (use .tex as suffix): quantcon.tex
You can store the variables and/or values used to split and fit in a file
Choose 1 to skip this step, 2 to store split and fit variables,
3 to store split variables and their values
Input your choice ([1:3], <cr>=1):
Input 2 to save fitted values and node IDs, 1 otherwise ([1:2], <cr>=2):
Input name of file to store node ID and fitted value of each case: quantcon.fit
Input 2 to write R function for predicting new cases, 1 otherwise ([1:2], <cr>=2):
Input file name: quantcon.r
Input rank of top variable to split root node ([1:431], <cr>=1):
Input file is created!
Run GUIDE with the command: guide < quantcon.in

Contents of quantcon.out

Quantile regression tree with quantile probability 0.9000
Pruning by cross-validation
Data description file: ce2021reg.dsc
Training sample file: ce2021.txt
Missing value code: NA
Records in data file start on line 2
Number of M variables associated with C variables: 19
384 N variables changed to S
D variable is INTRDVX
Piecewise constant model
Number of records in data file: 3965
Length of longest entry in data file: 11
Missing values found in D variable
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Warning: S variable DIRACC is constant
Warning: S variable TOTHVHRP is constant
Warning: S variable TOTHVHRC is constant
Warning: S variable ROTHWRFLC is constant
Warning: S variable WELFREBX is constant
Warning: S variable OTHLYRBX is constant
Warning: S variable OTHLNYRB is constant
Smallest and largest positive weights are  1.0725E+03 and  9.3902E+04

Wei-Yin Loh 117 GUIDE manual
7.1 Piecewise constant: one quantile  7  QUANTILE REGRESSION: CE DATA

Summary information for training sample of size 2487 (excluding observations with non-positive weight or missing values in d, e, t, r or z variables)
d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators),
s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
m=missing-value flag variable, p=periodic variable, w=weight
Levels of M variables are for missing values in associated variables

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 DIRACC</td>
<td>s</td>
<td>1.000</td>
<td>1.000</td>
<td>125</td>
</tr>
<tr>
<td>2 DIRACC_</td>
<td>m</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3 AGE_REF</td>
<td>s</td>
<td>19.00</td>
<td>87.00</td>
<td></td>
</tr>
<tr>
<td>4 AGE_REF_</td>
<td>m</td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>5 AGE2</td>
<td>s</td>
<td>21.00</td>
<td>87.00</td>
<td>1092</td>
</tr>
<tr>
<td>6 AGE2_</td>
<td>m</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>547 WHLFYR</td>
<td>c</td>
<td></td>
<td>1</td>
<td>2487</td>
</tr>
<tr>
<td>548 WHLFYR_</td>
<td>m</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>549 FFTAXDWE</td>
<td>s</td>
<td>-0.3368E+05</td>
<td>0.3380E+06</td>
<td></td>
</tr>
<tr>
<td>550 FSTAXDWE</td>
<td>s</td>
<td>-3074</td>
<td></td>
<td>0.5654E+05</td>
</tr>
</tbody>
</table>

Total #cases w/ #missing
#cases miss. D ord. vals #X-var #N-var #F-var #S-var
3965 1478 3965 1 0 0 384
#P-var #M-var #B-var #C-var #I-var
0 116 0 47 0

Number of cases used for training: 2487
Number of split variables: 431
Number of cases excluded due to 0 W or missing D variable: 1478

Constant fitted to cases with missing values in regressor variables
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE's for pruned tree: 0.2500

Weighted error estimates used for pruning
Warning: No interaction tests; too many predictor variables
Warning: All positive weights treated as 1
No nodewise interaction tests
Fraction of cases used for splitting each node: 1.0000
Maximum number of split levels: 12
Minimum node sample size: 24
Top-ranked variables and 1-df chi-squared values at root node
  1 0.6943E+02  CUTENURE
7.1 Piecewise constant: one quantile  7  QUANTILE REGRESSION: CE DATA

2 0.6324E+02  REFGEN
3 0.5982E+02  RENTEQVX
4 0.5957E+02  AGE_REF
5 0.5754E+02  AGE2

Size and CV Loss and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Nodes</th>
<th>Mean Loss</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median Loss</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>6.283E+07</td>
<td>4.914E+06</td>
<td>4.909E+06</td>
<td>6.446E+07</td>
<td>9.486E+06</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6.977E+07</td>
<td>5.683E+06</td>
<td>5.545E+06</td>
<td>6.830E+07</td>
<td>1.084E+07</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>7.116E+07</td>
<td>5.848E+06</td>
<td>5.732E+06</td>
<td>6.732E+06</td>
<td>1.353E+07</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>7.477E+07</td>
<td>5.750E+06</td>
<td>5.747E+06</td>
<td>6.854E+06</td>
<td>8.824E+06</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>9.107E+07</td>
<td>7.613E+06</td>
<td>5.017E+06</td>
<td>9.297E+07</td>
<td>5.472E+06</td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 9 terminal nodes
0-SE tree based on median is marked with + and has 8 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
** tree same as + tree
** tree same as -- tree

Following tree is based on mean CV with naive SE estimate (**)
7.1 Piecewise constant: one quantile

QUANTILE REGRESSION: CE DATA

26T 1311 1311 1 2.000E+03 RETSURVX
27T 24 24 1 4.000E+04 -
7T 252 252 1 2.000E+04 REF_RACE

Number of terminal nodes of final tree: 8
Total number of nodes of final tree: 15
Second best split variable (based on curvature test) at root node is REFGEN

Regression tree:
For categorical variable splits, values not in training data go to the right

Node 1: CUTENURE = "2", "5"
   Node 2: INC_RANK <= 0.81944155
      Node 4: INTRDVX sample quantile = 12000.000
      Node 2: INC_RANK > 0.81944155 or NA
         Node 5: RETPENPQ <= 90.250000
            Node 10: INTRDVX sample quantile = 141304.00
            Node 5: RETPENPQ > 90.250000 or NA
               Node 11: EENTMSCC <= 44.000000
                  Node 22: INTRDVX sample quantile = 25000.00
                  Node 23: INTRDVX sample quantile = 107121.00
               Node 2: INC_RANK /= "2", "5"
      Node 3: FFTAXOWE <= 30387.000
         Node 6: REF_RACE = "3"
            Node 12: INTRDVX sample quantile = 15000.000
            Node 6: REF_RACE /= "3"
               Node 13: STOCKYRX <= 83000.000 or NA
                  Node 26: INTRDVX sample quantile = 2000.000
                  Node 13: STOCKYRX > 83000.000
                     Node 27: INTRDVX sample quantile = 40000.000
         Node 3: FFTAXOWE > 30387.000 or NA
            Node 7: INTRDVX sample quantile = 20000.000

***********************************************************************
Predictor means below are weighted means of cases with no missing values.

WARNING: p-values below not adjusted for split search. For a bootstrap solution see:


Wei-Yin Loh
120
GUIDE manual
Figure 14 shows the quantile regression tree. The sample size (in italics) and 0.90-quantile are given beneath each terminal node.

### 7.2 Best simple linear

We demonstrate this with a linear 0.90-quantile regression tree.

#### 7.2.1 Input file creation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
Figure 14: GUIDE v.41.1 0.250-SE piecewise-constant 0.900-quantile regression tree for predicting INTRDVX. At each split, an observation goes to the left branch if and only if the condition is satisfied. Symbol ‘≤’ stands for ‘≤ or missing’. $S_1 = \{2, 5\}$. Sample size (in italics) and 0.900-quantile of INTRDVX printed below nodes. Terminal nodes with quantiles above and below value of 9800 at root node are colored yellow and orange respectively. Second best split variable at root node is REFGEN.
7.2 Best simple linear

Name of batch input file: quantlin.in
Input 1 for model fitting, 2 for importance or DIF scoring,
3 for data conversion ([1:3], <cr>=1):
Name of batch output file: quantlin.out
Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
Input 1 for classification, 2 for regression, 3 for propensity score grouping
Input your choice ([1:3], <cr>=1): 2
Choose type of regression model:
1=linear, 2=quantile, 3=Poisson, 4=censored response,
5=multiresponse or itemresponse, 6=longitudinal data (with T variables),
7=binary logistic regression.
Input choice ([1:7], <cr>=1): 2
Choose complexity of model to use at each node:
Choose 1 for multiple regression (recommended for prediction)
Choose 2 for best simple polynomial in one N or F variable
Choose 3 for constant fit (recommended for interpretability or if there is an R variable)
1: multiple linear, 2: best simple polynomial, 3: constant ([1:3], <cr>=3): 2
Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
Input quantile probability ([0.00:1.00], <cr>=0.50): 0.90
Input name of data description file (max 100 characters);
enclose with matching quotes if it has spaces: ce2021reg.dsc
Reading data description file ...
Training sample file: ce2021.txt
Missing value code: NA
Records in data file start on line 2
Number of M variables associated with C variables: 19
D variable is INTRDVX
Reading data file ...
Number of records in data file: 3965
Length of longest entry in data file: 11
Checking for missing values ...
Finished checking
Missing values found in D variable
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Finding number of levels of M variables associated with C variables ...
Assigning integer codes to values of 47 categorical variables
Finished assigning codes to 10 categorical variables
Finished assigning codes to 20 categorical variables
Finished assigning codes to 30 categorical variables
Finished assigning codes to 40 categorical variables
Associating missing values of N and S variables with M variable codes ...
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...

Wei-Yin Loh 123 GUIDE manual
Data checks complete
Creating missing value indicators ... 
Rereading data ...
Warning: N variable DIRACC is constant
Warning: N variable TOTHVHPR is constant
Warning: N variable TOTHVHRC is constant
Warning: N variable ROTHRFLEC is constant
Warning: N variable WELFREBX is constant
Warning: N variable OTHLYRBX is constant
Warning: N variable OTHLNYRB is constant
Smallest positive weight: 1.0725E+03
Largest positive weight: 9.3902E+04
Total #cases w/ #missing
  #cases   miss. D ord. vals #X-var #N-var #F-var #S-var
  3965     1478  3965        1   384    0    0
  #P-var   #M-var   #B-var #C-var #I-var
   0      116      0    47     0
Number of cases used for training: 2487
Number of split variables: 431
Number of cases excluded due to 0 W or missing D variable: 1478
Finished reading data file
Input 1 for unweighted, 2 for weighted error estimates during pruning ([1:2], <cr>=2):
Warning: No interaction tests; too many predictor variables
Warning: All positive weights treated as 1
Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):
Input file name to store LaTeX code (use .tex as suffix): quantlin.tex
You can store the variables and/or values used to split and fit in a file
Choose 1 to skip this step, 2 to store split and fit variables,
3 to store split variables and their values
Input your choice ([1:3], <cr>=1):
Input 2 to save regressor names in a file, 1 otherwise ([1:2], <cr>=2):
Input file name: quantlin.reg
Input 2 to save fitted values and node ID, 1 otherwise ([1:2], <cr>=2):
Input name of file to store node ID and fitted value of each case: quantlin.fit
Input 2 to write R function for predicting new cases, 1 otherwise ([1:2], <cr>=2):
Input file name: quantlin.r
Input rank of top variable to split root node ([1:431], <cr>=1):
Input file is created!
Run GUIDE with the command: guide < quantlin.in

Contents of quantlin1.out

Quantile regression tree with quantile probability 0.9000
No truncation of predicted values
Pruning by cross-validation

Wei-Yin Loh  124  GUIDE manual
Data description file: cereg.dsc
Training sample file: cedata.txt
Missing value code: NA
Records in data file start on line 2
Number of M variables associated with C variables: 33
D variable is INTRDVX
Piecewise simple linear or constant model
Powers are dropped if they are not significant at level 1.0000
Number of records in data file: 4693
Length of longest entry in data file: 11
Constant model fitted to incomplete cases in each node
Missing values found in D variable
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Warning: N variable OTHRINCB is constant
Warning: N variable NETRENTB is constant
Warning: N variable NETRNTBX is constant
Warning: N variable OTHLONBX is constant
Warning: N variable OTHLONB is constant
Smallest and largest positive weights are 1.3507E+03 and 7.0269E+04

Summary information for training sample of size 2922 (excluding observations with non-positive weight or missing values in d, e, t, r or z variables)

d=dependent, b=split and fit cat variable using indicator variables, c=split-only categorical, i=fit-only categorical (via indicators), s=split-only numerical, n=split and fit numerical, f=fit-only numerical, m=missing-value flag variable, p=periodic variable, w=weight
Levels of M variables are for missing values in associated variables

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DIRACC</td>
<td>c</td>
<td></td>
<td>2</td>
<td>116</td>
</tr>
<tr>
<td>2</td>
<td>DIRACC_</td>
<td>m</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>AGE_REF</td>
<td>n</td>
<td>18.00</td>
<td>87.00</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>AGE_REF_</td>
<td>m</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>AGE2</td>
<td>n</td>
<td>2.2000E+01</td>
<td>87.00</td>
<td>1225</td>
</tr>
<tr>
<td>6</td>
<td>AGE2_</td>
<td>m</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>:</td>
<td>FINLWT21</td>
<td>w</td>
<td>1351.</td>
<td>0.7027E+05</td>
<td></td>
</tr>
<tr>
<td>:</td>
<td>INTRDVX</td>
<td>d</td>
<td>1.000</td>
<td>0.9834E+05</td>
<td></td>
</tr>
<tr>
<td>:</td>
<td>FSTAXDWE</td>
<td>n</td>
<td>-2505.</td>
<td>0.5991E+05</td>
<td>0</td>
</tr>
<tr>
<td>652</td>
<td>FSTA_OWE</td>
<td>m</td>
<td></td>
<td>0.2782E+06</td>
<td></td>
</tr>
</tbody>
</table>

Wei-Yin Loh 125 GUIDE manual
### 7.2 Best simple linear

<table>
<thead>
<tr>
<th>Total #cases w/ #missing</th>
<th>#cases miss. D ord. vals</th>
<th>#X-var</th>
<th>#N-var</th>
<th>#F-var</th>
<th>#S-var</th>
</tr>
</thead>
<tbody>
<tr>
<td>4693</td>
<td>1771</td>
<td>4693</td>
<td>30</td>
<td>409</td>
<td>0</td>
</tr>
<tr>
<td>#P-var</td>
<td>#M-var</td>
<td>#B-var</td>
<td>#C-var</td>
<td>#I-var</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>168</td>
<td>0</td>
<td>44</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Number of cases used for training: 2922
Number of split variables: 453
Number of cases excluded due to 0 weight or missing D: 1771

Constant fitted to cases with missing values in regressor variables
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE’s for pruned tree: .2500

Weighted error estimates used for pruning
Warning: No interaction tests; too many predictor variables
Warning: All positive weights treated as 1
No nodewise interaction tests
Fraction of cases used for splitting each node: 1.0000
Maximum number of split levels: 12
Minimum node sample size: 30
Top-ranked variables and chi-squared values at root node
1 0.1258E+03 STATE
2 0.1148E+03 STOCKX
3 0.1010E+03 STOCKYRX

Size and CV Loss and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Tnodes</th>
<th>Mean Loss</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median Loss</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>71</td>
<td>4.341E+07</td>
<td>3.751E+06</td>
<td>5.517E+06</td>
<td>3.688E+07</td>
<td>7.638E+06</td>
</tr>
<tr>
<td>2</td>
<td>70</td>
<td>4.340E+07</td>
<td>3.751E+06</td>
<td>5.512E+06</td>
<td>3.688E+07</td>
<td>7.635E+06</td>
</tr>
<tr>
<td>44</td>
<td>14</td>
<td>4.435E+07</td>
<td>3.782E+06</td>
<td>5.659E+06</td>
<td>3.790E+07</td>
<td>7.155E+06</td>
</tr>
<tr>
<td>45**</td>
<td>13</td>
<td>4.179E+07</td>
<td>3.452E+06</td>
<td>5.504E+06</td>
<td>3.604E+07</td>
<td>5.284E+06</td>
</tr>
<tr>
<td>46--</td>
<td>11</td>
<td>4.296E+07</td>
<td>3.501E+06</td>
<td>5.728E+06</td>
<td>3.482E+07</td>
<td>7.137E+06</td>
</tr>
<tr>
<td>47++</td>
<td>10</td>
<td>4.384E+07</td>
<td>3.522E+06</td>
<td>5.917E+06</td>
<td>3.482E+07</td>
<td>7.626E+06</td>
</tr>
<tr>
<td>48</td>
<td>9</td>
<td>4.451E+07</td>
<td>3.544E+06</td>
<td>5.656E+06</td>
<td>3.956E+07</td>
<td>7.243E+06</td>
</tr>
<tr>
<td>49</td>
<td>8</td>
<td>4.872E+07</td>
<td>3.801E+06</td>
<td>5.118E+06</td>
<td>4.717E+07</td>
<td>6.062E+06</td>
</tr>
<tr>
<td>50</td>
<td>7</td>
<td>4.776E+07</td>
<td>3.582E+06</td>
<td>5.193E+06</td>
<td>4.391E+07</td>
<td>7.032E+06</td>
</tr>
<tr>
<td>51</td>
<td>4</td>
<td>4.931E+07</td>
<td>3.502E+06</td>
<td>3.598E+06</td>
<td>4.980E+07</td>
<td>4.804E+06</td>
</tr>
<tr>
<td>52</td>
<td>2</td>
<td>5.750E+07</td>
<td>4.270E+06</td>
<td>3.009E+06</td>
<td>5.606E+07</td>
<td>4.668E+06</td>
</tr>
</tbody>
</table>

Wei-Yin Loh 126 GUIDE manual
7.2 Best simple linear  

<p>| Cases fit Give the number of cases used to fit node |
|---------------------------------|-----------------|---------------------------|</p>
<table>
<thead>
<tr>
<th>Label</th>
<th>Total cases</th>
<th>Cases Matrix</th>
<th>Node</th>
<th>Split</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2922</td>
<td>2922</td>
<td>2</td>
<td>9.500E+03</td>
<td>STATE</td>
</tr>
<tr>
<td>2T</td>
<td>250</td>
<td>250</td>
<td>2</td>
<td>9.834E+04</td>
<td>PERSUT64</td>
</tr>
<tr>
<td>3</td>
<td>2672</td>
<td>2672</td>
<td>2</td>
<td>6.120E+03</td>
<td>STOCKX</td>
</tr>
<tr>
<td>6</td>
<td>2540</td>
<td>2540</td>
<td>2</td>
<td>5.000E+03</td>
<td>FINCATAX</td>
</tr>
<tr>
<td>12</td>
<td>1658</td>
<td>1658</td>
<td>2</td>
<td>3.400E+03</td>
<td>PERSUT64</td>
</tr>
<tr>
<td>24T</td>
<td>936</td>
<td>936</td>
<td>2</td>
<td>1.200E+03</td>
<td>AGE_REF</td>
</tr>
<tr>
<td>25</td>
<td>722</td>
<td>722</td>
<td>2</td>
<td>8.000E+03</td>
<td>TOTXEST</td>
</tr>
<tr>
<td>50T</td>
<td>349</td>
<td>349</td>
<td>2</td>
<td>3.600E+03</td>
<td>SLOCTAXX</td>
</tr>
<tr>
<td>51</td>
<td>373</td>
<td>373</td>
<td>2</td>
<td>1.200E+04</td>
<td>PSU</td>
</tr>
<tr>
<td>102T</td>
<td>36</td>
<td>36</td>
<td>2</td>
<td>2.100E+04</td>
<td>-</td>
</tr>
<tr>
<td>103T</td>
<td>337</td>
<td>337</td>
<td>2</td>
<td>9.500E+03</td>
<td>BUILDING</td>
</tr>
<tr>
<td>13</td>
<td>882</td>
<td>882</td>
<td>2</td>
<td>1.000E+04</td>
<td>INCLASS2</td>
</tr>
<tr>
<td>26</td>
<td>89</td>
<td>89</td>
<td>2</td>
<td>9.834E+04</td>
<td>FSALARYX</td>
</tr>
<tr>
<td>52T</td>
<td>34</td>
<td>34</td>
<td>2</td>
<td>9.834E+04</td>
<td>-</td>
</tr>
<tr>
<td>53T</td>
<td>55</td>
<td>55</td>
<td>2</td>
<td>1.000E+03</td>
<td>-</td>
</tr>
<tr>
<td>27</td>
<td>793</td>
<td>793</td>
<td>2</td>
<td>9.000E+03</td>
<td>CUTEENURE</td>
</tr>
<tr>
<td>54</td>
<td>226</td>
<td>226</td>
<td>2</td>
<td>2.206E+04</td>
<td>FEDTAXX</td>
</tr>
<tr>
<td>108</td>
<td>194</td>
<td>194</td>
<td>2</td>
<td>1.500E+04</td>
<td>FJSSDEDX</td>
</tr>
<tr>
<td>216T</td>
<td>46</td>
<td>46</td>
<td>2</td>
<td>2.370E+04</td>
<td>-</td>
</tr>
<tr>
<td>217T</td>
<td>148</td>
<td>148</td>
<td>2</td>
<td>9.000E+03</td>
<td>TUTEXPPQ</td>
</tr>
<tr>
<td>109T</td>
<td>32</td>
<td>32</td>
<td>2</td>
<td>9.834E+04</td>
<td>-</td>
</tr>
<tr>
<td>55T</td>
<td>567</td>
<td>567</td>
<td>2</td>
<td>2.000E+03</td>
<td>FEDTAXX</td>
</tr>
<tr>
<td>7</td>
<td>132</td>
<td>132</td>
<td>2</td>
<td>9.834E+04</td>
<td>STOCKX</td>
</tr>
<tr>
<td>14T</td>
<td>102</td>
<td>102</td>
<td>2</td>
<td>2.000E+04</td>
<td>FEDTAXX</td>
</tr>
<tr>
<td>15T</td>
<td>30</td>
<td>30</td>
<td>2</td>
<td>9.834E+04</td>
<td>-</td>
</tr>
</tbody>
</table>

Number of terminal nodes of final tree: 13  
Total number of nodes of final tree: 25
7.2  Best simple linear

Second best split variable (based on curvature test) at root node is STOCKX

Regression tree:
For categorical variable splits, values not in training data go to the right

Node 1: STATE = "8", "11", "23", "32", "34", "45", "53", "54"
Node 2: INTRDVX sample quantile = 98338.000
Node 1: STATE /= "8", "11", "23", "32", "34", "45", "53", "54"
Node 3: STOCKX <= 86500.000 or STOCKX = NA & STOCKX_ = "A"
   Node 6: FINCATAx <= 98231.000
      Node 12: PERSOT64 <= .5000000
         Node 24: INTRDVX sample quantile = 1200.0000
      Node 12: PERSOT64 > .5000000 or NA
         Node 25: TOTXEST <= 82.500000
            Node 50: INTRDVX sample quantile = 3600.0000
         Node 25: TOTXEST > 82.500000 or NA
            Node 51: PSU = "1110", "1207", "1210", "1316", "1318", "1319", "1320"
               Node 102: INTRDVX sample quantile = 21000.000
            Node 51: PSU /= "1110", "1207", "1210", "1316", "1318", "1319", "1320"
               Node 103: INTRDVX sample quantile = 9500.000
      Node 6: FINCATAx > 98231.000 or NA
      Node 13: INCLASS2 <= 5.5000000
         Node 26: FSALARYX <= 41500.000
            Node 52: INTRDVX sample quantile = 98338.000
         Node 26: FSALARYX > 41500.000 or NA
            Node 53: INTRDVX sample quantile = 1000.000
      Node 13: INCLASS2 > 5.5000000 or NA
      Node 27: CUTENURE = "2"
         Node 54: FEDTAXX <= 3637.0000 or NA
            Node 108: FJSSDEDX <= 6440.0000
               Node 216: INTRDVX sample quantile = 23700.000
            Node 108: FJSSDEDX > 6440.0000 or NA
               Node 217: INTRDVX sample quantile = 9000.000
         Node 54: FEDTAXX > 3637.0000
            Node 109: INTRDVX sample quantile = 98338.000
      Node 27: CUTENURE /= "2"
         Node 55: INTRDVX sample quantile = 2000.0000
  Node 3: not (STOCKX <= 86500.000 or STOCKX = NA & STOCKX_ = "A")
  Node 7: STOCKX <= 478846.50 or STOCKX = NA & STOCKX_ = "C"
     Node 14: INTRDVX sample quantile = 20000.000
     Node 7: not (STOCKX <= 478846.50 or STOCKX = NA & STOCKX_ = "C")
        Node 15: INTRDVX sample quantile = 98338.000

******************************************************************************

Predictor means below are weighted means of cases with no missing values.
Regression coefficients are computed from the complete cases.

WARNING: p-values below not adjusted for split search. For a bootstrap solution see:


Node 1: Intermediate node
A case goes into Node 2 if STATE = "8", "11", "23", "32", "34", "45", "53", "54"
STATE mode = "NA"
Coefficients of quantile regression function:
Regressor Coefficient Minimum Mean Maximum
Constant -6539.
AGE_REF 297.7 18.00 55.40 87.00
If regressors have missing values, predicted quantile = 9500.00
----------------------------
Node 2: Terminal node
Coefficients of quantile regression function:
Regressor Coefficient Minimum Mean Maximum
Constant 1114.
FINCATAx 0.3297 -0.1374E+05 0.1119E+06 0.8418E+06
If regressors have missing values, predicted quantile = 98338.0
----------------------------
Node 3: Intermediate node
A case goes into Node 6 if STOCKX <= 86500.000 or STOCKX_ = "A"
STOCKX mean = 404023.36
----------------------------
Node 7: Intermediate node
A case goes into Node 14 if STOCKX <= 478846.50 or STOCKX_ = "C"
STOCKX mean = 1195543.9
----------------------------
Node 14: Terminal node
Coefficients of quantile regression function:
Regressor Coefficient Minimum Mean Maximum
Constant 0.2000E+05
TEXTILCQ 432.8 0.000 9.776 517.0
If regressors have missing values, predicted quantile = 20000.0
----------------------------
7.3 Two quantiles: checking variance heterogeneity

Checking variance homogeneity in the residuals is a standard practice in fitting regression models. Here we show how GUIDE can do this by constructing a quantile regression tree models for the 25th and 75th quantiles simultaneously.

7.3.1 Input file creation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: twoquant.in
   Input 1 for model fitting, 2 for importance or DIF scoring,
   3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: twoquant.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1): 2
   Choose type of regression model:
   1=linear, 2=quantile, 3=Poisson, 4=censored response,
   5=multiresponse or itemresponse, 6=longitudinal data (with T variables),
   7=binary logistic regression.
   Input choice ([1:7], <cr>=1): 2
   Choose complexity of model to use at each node:
   Choose 1 for multiple regression (recommended for prediction)
   Choose 2 for best simple polynomial in one N or F variable
   Choose 3 for constant fit (recommended for interpretability or if there is an R variable)
   1: multiple linear, 2: best simple polynomial, 3: constant ([1:3], <cr>=3):
   Input 1 for default options, 2 otherwise ([1:2], <cr>=1):

Figure 15 shows the 0.90-quantile regression tree.
Figure 15: GUIDE v.40.2 0.250-SE piecewise simple linear 0.900-quantile regression tree (constant fitted to incomplete cases) for predicting INTRDVX. At each split, an observation goes to the left branch if and only if the condition is satisfied. Symbol ‘≤∗’ stands for ‘≤ or missing’. \(S_1 = \{8, 11, 23, 32, 34, 45, 53, 54\}\). \(S_2 = \{1110, 1207, 1210, 1316, 1318, 1319, 1320\}\). Sample size (in italics), 0.900-quantile of INTRDVX, and sign and name of best regressor printed below nodes. Terminal nodes with quantiles above and below value of 9500 at root node are colored yellow and orange respectively. Second best split variable at root node is \(STOCKX\).
7.3 Two quantiles

Input 1 for 1 quantile, 2 for 2 quantiles ([1:2], \(<cr>=1\)): 2
Input 1st quantile probability ([0.00:1.00], \(<cr>=0.25\)):
Input 2nd quantile probability ([0.00:1.00], \(<cr>=0.75\)):
Input name of data description file (max 100 characters);
enclose with matching quotes if it has spaces: ce2021reg.dsc
Reading data description file ...
Training sample file: ce2021.txt
Missing value code: NA
Records in data file start on line 2
Number of M variables associated with C variables: 19
384 N variables changed to S
D variable is INTRDVX
Reading data file ...
Number of records in data file: 3965
Length of longest entry in data file: 11
Checking for missing values ...
Finished checking
Missing values found in D variable
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Finding number of levels of M variables associated with C variables ...
Assigning integer codes to values of 47 categorical variables
Finished assigning codes to 10 categorical variables
Finished assigning codes to 20 categorical variables
Finished assigning codes to 30 categorical variables
Finished assigning codes to 40 categorical variables
Associating missing values of N and S variables with M variable codes ...
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...
Data checks complete
Creating missing value indicators ...
Rereading data ...
Warning: S variable DIRACC is constant
Warning: S variable TUTHVHRP is constant
Warning: S variable TOTHVHRC is constant
Warning: S variable ROTHRLFCL is constant
Warning: S variable WELFREBX is constant
Warning: S variable OTHLYRBX is constant
Warning: S variable OTHLNYRB is constant
Smallest positive weight: 1.0725E+03
Largest positive weight: 9.3902E+04

<table>
<thead>
<tr>
<th>Total</th>
<th>#cases w/ #missing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#cases miss. D ord. vals #X-var #N-var #F-var #S-var</td>
</tr>
<tr>
<td>3965</td>
<td>1478 3965 1 0 0 384</td>
</tr>
</tbody>
</table>

Wei-Yin Loh 132 GUIDE manual
7.3 Two quantiles

<table>
<thead>
<tr>
<th>#P-var</th>
<th>#M-var</th>
<th>#B-var</th>
<th>#C-var</th>
<th>#I-var</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>116</td>
<td>0</td>
<td>47</td>
<td>0</td>
</tr>
</tbody>
</table>

Number of cases used for training: 2487
Number of split variables: 431
Number of cases excluded due to 0 W or missing D variable: 1478
Finished reading data file

Input 1 for unweighted, 2 for weighted error estimates during pruning ([1:2], <cr>=2):
Warning: No interaction tests; too many predictor variables
Warning: All positive weights treated as 1
Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):
Input file name to store LaTeX code (use .tex as suffix): twoquant.tex
You can store the variables and/or values used to split and fit in a file
Choose 1 to skip this step, 2 to store split and fit variables, 3 to store split variables and their values
Input your choice ([1:3], <cr>=1):
Input 2 to save fitted values and node IDs, 1 otherwise ([1:2], <cr>=2):
Input name of file to store node ID and fitted value of each case: twoquant.fit
Input 2 to write R function for predicting new cases, 1 otherwise ([1:2], <cr>=2):
Input file name: twoquant.r
Input rank of top variable to split root node ([1:431], <cr>=1):
Input file is created!
Run GUIDE with the command: guide < twoquant.in

7.3.2 Output file

Dual-quantile regression tree with 0.2500 and 0.7500 quantiles
Pruning by cross-validation
Data description file: ce2021reg.dsc
Training sample file: ce2021.txt
Missing value code: NA
Records in data file start on line 2
Number of M variables associated with C variables: 19
384 N variables changed to S
D variable is INTRDVX
Piecewise constant model
Number of records in data file: 3965
Length of longest entry in data file: 11
Missing values found in D variable
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Warning: S variable DIRACC is constant
Warning: S variable TOTHVHRP is constant
Warning: S variable TOTHVHRC is constant
Warning: S variable ROTHFLC is constant

Wei-Yin Loh

GUIDE manual
7.3 Two quantiles

Summary information for training sample of size 2487 (excluding observations with non-positive weight or missing values in d, e, t, r or z variables)

d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators),
s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
m=missing-value flag variable, p=periodic variable, w=weight

Levels of M variables are for missing values in associated variables

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DIRACC</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td>125</td>
</tr>
<tr>
<td>2</td>
<td>DIRACC</td>
<td>m</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>AGE_REF</td>
<td>s</td>
<td>19.00</td>
<td>87.00</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>AGE_REF</td>
<td>m</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>AGE2</td>
<td>s</td>
<td>21.00</td>
<td>87.00</td>
<td>1092</td>
</tr>
<tr>
<td>6</td>
<td>AGE2</td>
<td>m</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total</th>
<th>#cases w/</th>
<th>#missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>#cases</td>
<td>3965</td>
<td>1478</td>
</tr>
<tr>
<td>miss. D</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ord. vals</td>
<td>3965</td>
<td>1</td>
</tr>
<tr>
<td>#A-var</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>#N-var</td>
<td>0</td>
<td>384</td>
</tr>
<tr>
<td>#F-var</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>#S-var</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>#P-var</td>
<td>0</td>
<td>116</td>
</tr>
<tr>
<td>#B-var</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>#C-var</td>
<td>47</td>
<td>0</td>
</tr>
<tr>
<td>#I-var</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Number of cases used for training: 2487
Number of split variables: 431
Number of cases excluded due to 0 W or missing D variable: 1478

Constant fitted to cases with missing values in regressor variables
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE's for pruned tree: 0.2500

Weighted error estimates used for pruning
Warning: No interaction tests; too many predictor variables
Warning: All positive weights treated as 1
No nodewise interaction tests
Fraction of cases used for splitting each node: 1.0000
Maximum number of split levels: 12
Minimum node sample size: 24
7.3 Two quantiles

Top-ranked variables and 1-df chi-squared values at root node

1 0.1744E+03 STATE
2 0.1192E+03 FINCBTAX
3 0.1135E+03 INC_RANK
4 0.9547E+02 OCCUCOD1

Size and CV Loss and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Nodes</th>
<th>Mean Loss</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median Loss</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>76</td>
<td>1.016E+08</td>
<td>7.031E+06</td>
<td>5.504E+06</td>
<td>1.017E+08</td>
<td>6.605E+06</td>
</tr>
<tr>
<td>2</td>
<td>75</td>
<td>1.016E+08</td>
<td>7.031E+06</td>
<td>5.504E+06</td>
<td>1.017E+08</td>
<td>6.605E+06</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>42</td>
<td>15</td>
<td>1.005E+08</td>
<td>7.073E+06</td>
<td>5.612E+06</td>
<td>9.975E+07</td>
<td>6.599E+06</td>
</tr>
<tr>
<td>43</td>
<td>13</td>
<td>1.005E+08</td>
<td>7.144E+06</td>
<td>5.667E+06</td>
<td>9.975E+07</td>
<td>6.463E+06</td>
</tr>
<tr>
<td>44*</td>
<td>7</td>
<td>1.003E+08</td>
<td>7.152E+06</td>
<td>5.731E+06</td>
<td>9.967E+07</td>
<td>6.602E+06</td>
</tr>
<tr>
<td>45++</td>
<td>6</td>
<td>1.004E+08</td>
<td>7.153E+06</td>
<td>5.709E+06</td>
<td>9.967E+07</td>
<td>6.393E+06</td>
</tr>
<tr>
<td>46**</td>
<td>5</td>
<td>1.010E+08</td>
<td>7.361E+06</td>
<td>5.788E+06</td>
<td>1.020E+08</td>
<td>6.408E+06</td>
</tr>
<tr>
<td>47</td>
<td>1</td>
<td>1.164E+08</td>
<td>9.292E+06</td>
<td>5.964E+06</td>
<td>1.190E+08</td>
<td>7.350E+06</td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 7 terminal nodes
0-SE tree based on median is marked with + and has 6 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
** tree same as -- tree
+ tree same as ++ tree

Following tree is based on mean CV with naive SE estimate (**)
Number of terminal nodes of final tree: 5
Total number of nodes of final tree: 9
Second best split variable (based on curvature test) at root node is FINCBTAX

Regression tree:
For categorical variable splits, values not in training data go to the right

Node 1: STATE = "19", "24", "31", "40", "45", "49", "51", "54"
  Node 2: INTRDVX sample quantiles = 5.000000, 250.0000
Node 1: STATE /= "19", "24", "31", "40", "45", "49", "51", "54"
  Node 3: INC_RANK /= 0.88093190
  Node 6: INTRDVX sample quantiles = 15.000000, 1500.0000
  Node 3: INC_RANK > 0.88093190 or NA
  Node 7: INCNONW2 = "1"
    or (INCNONW2 = NA & INCN_NW2 = "A")
  Node 14: INCNONW2 = "1"
    Node 28: INTRDVX sample quantiles = 2191.0000, 107121.00
  Node 14: INCNONW2 /= "1"
    Node 29: INTRDVX sample quantiles = 100.0000, 3000.0000
  Node 7: INCNONW2 /= "1"
    & not (INCNONW2 = NA & INCN_NW2 = "A")
  Node 15: INTRDVX sample quantiles = 200.000000, 20000.00

***************************************************************
Predictor means below are weighted means of cases with no missing values.
WARNING: p-values below not adjusted for split search. For a bootstrap solution see:

Node 1: Intermediate node
A case goes into Node 2 if STATE = "19", "24", "31", "40", "45", "49", "51", "54"
  STATE mode = "6"
  Sample 0.250-quantile, 0.750-quantile, and median:
  1.5000E+01  1.9300E+03  1.7400E+02

Node 2: Terminal node
Sample 0.250-quantile, 0.750-quantile, and median:
7.3 Two quantiles

5.0000E+00 2.5000E+02 2.0000E+01

---------------------------------

Node 3: Intermediate node
A case goes into Node 6 if INC_RANK <= 0.88093190
INC_RANK mean = 0.65419245

---------------------------------

Node 6: Terminal node
Sample 0.250-quantile, 0.750-quantile, and median:
1.5000E+01 1.5000E+03 1.5000E+02

---------------------------------

Node 7: Intermediate node
A case goes into Node 14 if INCNONW2 = "1"
or INCNONW2 = NA & INCN_NW2 = "A"
INCN_NW2 mode = "A"

---------------------------------

Node 14: Intermediate node
A case goes into Node 28 if INCNONW2 = "1"
INCN_NW2 mode = "A"

---------------------------------

Node 28: Terminal node
Sample 0.250-quantile, 0.750-quantile, and median:
2.1910E+03 1.0712E+05 2.5879E+04

---------------------------------

Node 29: Terminal node
Sample 0.250-quantile, 0.750-quantile, and median:
1.0000E+02 3.0000E+03 5.0000E+02

---------------------------------

Node 15: Terminal node
Sample 0.250-quantile, 0.750-quantile, and median:
2.0000E+02 2.0000E+04 6.5000E+03

---------------------------------

Observed and fitted values are stored in twoquant.fit
LaTeX code for tree is in twoquant.tex
R code is stored in twoquant.r

Figure 16 shows the tree. Beneath each terminal node are three numbers. The first (in italics) is the node sample size. The other two are the sample 0.75 and 0.25-quantiles in the node. The large between-node variations in the inter-quartile ranges in the nodes indicates substantial variance heterogeneity.
Figure 16: GUIDE v.41.1 0.250-SE piecewise-constant 0.250 and 0.750-quantile regression tree for predicting INTRDVX. At each split, an observation goes to the left branch if and only if the condition is satisfied. $S_1 = \{19, 24, 31, 40, 45, 49, 51, 54\}$. $S_2 = \{1\}$; $S_2^* = \{A\}$. Sample size (in italics) and sample 0.750 and 0.250-quantiles of INTRDVX printed below nodes. Second best split variable at root node is FINCBTAX.

Wei-Yin Loh 138 GUIDE manual
8 Periodic variables: NHTSA data

Periodic variables that have a cyclic property, such as angular measurements, hour of day, day of week, and month of year, can be designated as \( P \) variables in the description file. There can be multiple \( P \) variables in the same data set. Unlike the other types of variables, each line in the description file containing a \( P \) variable must have the value of its period (e.g., 360 for angular measurements, 24 for hour of day, 7 for day of week, and 12 for month of year) immediately after \( P \) on the same line. GUIDE does not allow \( P \) variables to have missing-value flag (M) variables.

The National Highway Traffic Safety Administration (NHTSA) has been conducting vehicle crash tests since 1972. Data from 3310 crash tests are in the file nhtsadatam.txt (see www.nrd.nhtsa.dot.gov/database/veh/ for more information). Variable HIC (head injury criterion) is a measure of severity of head injury. Experts believe that HIC > 999 is life threatening. Table 9 gives the definitions of the variables appearing in the models below. Besides missing values, there are many variables with illogical values (such as negative values for diameter). To identify these values, we adopt the strategy in the CE data of creating a missing-value flag variable for each variable having illogical values, with the flags being A, B, and D for validly missing, illogical, and valid response, respectively. The data also contain some angular variables (with periods of 360 degrees and for which 0 degrees indicates straight-ahead or head-on) that are defined as \( P \) in the description file nhtsadsc.txt below.

nhtsadatam.txt
NA
2
1 BARRIG c
2 BARSHP b
3 BARANG p 360
4 BARDIA n
5 OCCWT n
6 OCCWT_m
7 DUMSIZ c
8 HH n
9 HH_m
10 HW n
11 HW_m
12 HR n
13 HR_m
14 HS n
15 HS_m
16 CD n
17 CD_ m
18 CS n
19 CS_ m
20 AD n
21 AD_ m
22 HD n
23 HD_ m
24 KD n
25 KD_ m
26 HB n
27 HB_ m
28 NB n
29 NB_ m
30 CB n
31 CB_ m
32 KB n
33 SEPOSN c
34 HIC d
35 TKSURF c
36 TKCOND c
37 CLSSPD n
38 CLSSPD_ m
39 IMPANG p 360
40 OFFSET n
41 IMPPNT n
42 MADE c
43 MODEL c
44 YEAR n
45 BODY c
46 ENGINE c
47 ENGDSP n
48 ENGDSP_ m
49 TRANSM c
50 VEHTWT n
51 VEHTWT_ m
52 CURBWT n
53 WHLBAS n
54 WHLBAS_ m
55 VEHLEN n
56 VEHLEN_ m
57 VEHWID n
58 VEHWID_ m
59 VEHCG n
60 VEHCG_ m
61 COLMEL c
62 BXI n
63 BX1_  m
64 BX2_  n
65 BX2_  m
66 BX3_  n
67 BX3_  m
68 BX4_  n
69 BX4_  m
70 BX5_  n
71 BX5_  m
72 BX6_  n
73 BX6_  m
74 BX7_  n
75 BX7_  m
76 BX8_  n
77 BX8_  m
78 BX9_  n
79 BX9_  m
80 BX10_ n
81 BX10_ m
82 BX11_ n
83 BX11_ m
84 BX12_ n
85 BX12_ m
86 BX13_ n
87 BX13_ m
88 BX14_ n
89 BX14_ m
90 BX15_ n
91 BX15_ m
92 BX16_ n
93 BX16_ m
94 BX17_ n
95 BX17_ m
96 BX18_ n
97 BX18_ m
98 BX19_ n
99 BX19_ m
100 BX20_ n
101 BX20_ m
102 BX21_ n
103 BX21_ m
104 VEHSPD_ n
105 VEHSPD_ m
106 CRBANG p 360
107 PDDF p 360
108 CARANG p 360
Table 9: Some variable definitions for NHTSA data

<table>
<thead>
<tr>
<th>Variable</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>BARSHP</td>
<td>barrier shape (21 values)</td>
</tr>
<tr>
<td>BX2</td>
<td>distance from rear surface of vehicle to front of engine (mm)</td>
</tr>
<tr>
<td>BX5</td>
<td>distance from rear surface of vehicle to upper leading edge of left door (mm)</td>
</tr>
<tr>
<td>BX8</td>
<td>distance from rear surface of vehicle to upper trailing edge of right door (mm)</td>
</tr>
<tr>
<td>BX12</td>
<td>distance from rear surface of vehicle to bottom of a post of right side (mm)</td>
</tr>
<tr>
<td>COLMEC</td>
<td>steering column collapse mechanism (9 values)</td>
</tr>
<tr>
<td>ENGDSP</td>
<td>engine displacement (liters)</td>
</tr>
<tr>
<td>IMPANG</td>
<td>impact angle (clockwise with 0 degrees being straight ahead)</td>
</tr>
<tr>
<td>OCCAGE</td>
<td>dummy occupant age</td>
</tr>
<tr>
<td>PDOF</td>
<td>principal direction of force (degrees)</td>
</tr>
<tr>
<td>TRANSM</td>
<td>transmission type (9 values)</td>
</tr>
<tr>
<td>VEHTWT</td>
<td>vehicle test weight (kg)</td>
</tr>
<tr>
<td>VEHSPD</td>
<td>vehicle speed (km/h)</td>
</tr>
<tr>
<td>VEHWID</td>
<td>vehicle width (mm)</td>
</tr>
<tr>
<td>WHLBS</td>
<td>wheel base (mm)</td>
</tr>
<tr>
<td>YEAR</td>
<td>vehicle model year (1972–2017)</td>
</tr>
</tbody>
</table>

We show the results of fitting a piecewise-linear regression tree here.

8.1 Input file creation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: lin.in
   Input 1 for model fitting, 2 for importance or DIF scoring, 3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: lin.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1): 2
   Choose type of regression model:
8.1 Input file creation

1=linear, 2=quantile, 3=Poisson, 4=censored response, 5=multiresponse or itemresponse, 6=longitudinal data (with T variables), 7=binary logistic regression.

Input choice ([1:7], <cr>=1):
Input 1 for least squares, 2 least median of squares ([1:2], <cr>=1):

Choose complexity of model to use at each node:
Choose 0 for stepwise linear regression (recommended for prediction)
Choose 1 for multiple regression
Choose 2 for best simple polynomial in one N or F variable
Choose 3 for constant fit (recommended for interpretability or if there is an R variable)

0: stepwise linear, 1: multiple linear, 2: best simple polynomial, 3: constant, 4: best simple stepwise ANCOVA ([0:4], <cr>=3): 2
Input 1 for default options, 2 otherwise ([1:2], <cr>=1):

Input name of data description file (max 100 characters); enclose with matching quotes if it has spaces: nhtsadsc.txt

Reading data description file ...

Training sample file: nhtsadatam.txt
Missing value code: NA
Records in data file start on line 2
Warning: B variables changed to C
D variable is HIC
Reading data file ...
Number of records in data file: 3310
Length of longest entry in data file: 19
Checking for missing values ...
Finished checking
Missing values found in D variable
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Assigning integer codes to values of 13 categorical variables
Finished assigning codes to 10 categorical variables
Associating missing values of N, P and S variables with M variable codes ...
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...
Data checks complete
Creating missing value indicators ...
Rereading data ...

Total #cases w/ #missing
#cases miss. D ord. vals #X-var #N-var #F-var #S-var
3310 34 3310 2 48 0 0

#P-var #M-var #B-var #C-var #I-var
6 42 0 13 0

No weight variable in data file
Number of cases used for training: 3276
8.2 Results

Least squares regression tree
Predictions truncated at global min. and max. of D sample values
Pruning by cross-validation
Data description file: nhtsadsc.txt
Training sample file: nhtsadatam.txt
Missing value code: NA
Records in data file start on line 2
Warning: B variables changed to C
D variable is HIC
Piecewise simple linear or constant model
Powers are dropped if they are not significant at level 0.0500
Number of records in data file: 3310
Length of longest entry in data file: 19
Missing values found in D variable
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables

Summary information for training sample of size 3276 (excluding observations with non-positive weight or missing values in d, e, t, r or z variables)
d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators),
s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
### 8.2 Results

`m`=missing-value flag variable, `p`=periodic variable, `w`=weight

Levels of `m` variables are for missing values in associated variables.

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BARRIG</td>
<td>c</td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>BARSHP</td>
<td>c</td>
<td></td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>BARANG</td>
<td>p</td>
<td>0.000</td>
<td>330.0</td>
<td>360</td>
</tr>
<tr>
<td>4</td>
<td>BARDIA</td>
<td>n</td>
<td>1.910E+02</td>
<td>1000.</td>
<td>2807</td>
</tr>
<tr>
<td>5</td>
<td>OCCWT</td>
<td>n</td>
<td>7.2000E+01</td>
<td>83.00</td>
<td>3265</td>
</tr>
<tr>
<td>6</td>
<td>OCCWT_</td>
<td>m</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>106</td>
<td>CRBANG</td>
<td>p</td>
<td>0.000</td>
<td>315.0</td>
<td>360</td>
</tr>
<tr>
<td>107</td>
<td>PDOF</td>
<td>p</td>
<td>0.000</td>
<td>345.0</td>
<td>360</td>
</tr>
<tr>
<td>108</td>
<td>CARANG</td>
<td>p</td>
<td>0.000</td>
<td>99.00</td>
<td>360</td>
</tr>
<tr>
<td>109</td>
<td>VEHOR</td>
<td>p</td>
<td>0.000</td>
<td>90.00</td>
<td>360</td>
</tr>
<tr>
<td>110</td>
<td>RSTFRT</td>
<td>c</td>
<td></td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

Total #cases w/ #missing

<table>
<thead>
<tr>
<th>#cases</th>
<th>miss.</th>
<th>D ord. vals</th>
<th>#X-var</th>
<th>#N-var</th>
<th>#F-var</th>
<th>#S-var</th>
<th>#P-var</th>
<th>#M-var</th>
<th>#B-var</th>
<th>#C-var</th>
<th>#I-var</th>
</tr>
</thead>
<tbody>
<tr>
<td>3310</td>
<td>34</td>
<td>3310</td>
<td>2</td>
<td>48</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>42</td>
<td>0</td>
<td>13</td>
<td>0</td>
</tr>
</tbody>
</table>

No weight variable in data file.

Number of cases used for training: 3276

Number of split variables: 61

Number of cases excluded due to 0 W or missing D variable: 34

Constant fitted to cases with missing values in regressor variables.

Pruning by v-fold cross-validation, with `v` = 10

Selected tree is based on mean of CV estimates.

Number of SE's for pruned tree: 0.2500

Nodewise interaction tests on all variables.

Fraction of cases used for splitting each node: 1.0000

Maximum number of split levels: 13

Minimum node sample size: 33

Top-ranked variables and 1-df chi-squared values at root node:

<table>
<thead>
<tr>
<th>Rank</th>
<th>Chi-Square</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2525E+03</td>
<td>BARSHP</td>
</tr>
<tr>
<td>2</td>
<td>0.1423E+03</td>
<td>IMPANG</td>
</tr>
<tr>
<td>3</td>
<td>0.1245E+03</td>
<td>BARDIA</td>
</tr>
<tr>
<td>4</td>
<td>0.1245E+03</td>
<td>BODY</td>
</tr>
</tbody>
</table>
8.2 Results

PERIODIC VARIABLES: NHTSA DATA

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.1204E+03</td>
</tr>
<tr>
<td>66</td>
<td>0.1448E+01</td>
</tr>
<tr>
<td>67</td>
<td>0.6191E-02</td>
</tr>
</tbody>
</table>

Size and CV MSE and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Nodes</th>
<th>Mean MSE</th>
<th>SE(Mean)</th>
<th>Median MSE</th>
<th>BSE(Mean)</th>
<th>Median MSE</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>60</td>
<td>3.748E+05</td>
<td>6.535E+04</td>
<td>7.046E+04</td>
<td>3.413E+05</td>
<td>5.078E+04</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>58</td>
<td>3.748E+05</td>
<td>6.535E+04</td>
<td>7.046E+04</td>
<td>3.413E+05</td>
<td>5.078E+04</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>56</td>
<td>3.748E+05</td>
<td>6.535E+04</td>
<td>7.046E+04</td>
<td>3.413E+05</td>
<td>5.078E+04</td>
<td></td>
</tr>
<tr>
<td>23+</td>
<td>28</td>
<td>3.640E+05</td>
<td>6.529E+04</td>
<td>6.996E+04</td>
<td>3.308E+05</td>
<td>4.855E+04</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>21</td>
<td>3.676E+05</td>
<td>6.605E+04</td>
<td>7.105E+04</td>
<td>3.425E+05</td>
<td>5.089E+04</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>20</td>
<td>3.673E+05</td>
<td>6.605E+04</td>
<td>7.105E+04</td>
<td>3.425E+05</td>
<td>5.067E+04</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>19</td>
<td>3.711E+05</td>
<td>6.644E+04</td>
<td>7.118E+04</td>
<td>3.398E+05</td>
<td>5.463E+04</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>17</td>
<td>3.686E+05</td>
<td>6.626E+04</td>
<td>7.117E+04</td>
<td>3.398E+05</td>
<td>5.222E+04</td>
<td></td>
</tr>
<tr>
<td>28++</td>
<td>16</td>
<td>3.691E+05</td>
<td>6.632E+04</td>
<td>7.125E+04</td>
<td>3.428E+05</td>
<td>5.255E+04</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>15</td>
<td>3.640E+05</td>
<td>6.635E+04</td>
<td>6.886E+04</td>
<td>3.572E+05</td>
<td>5.342E+04</td>
<td></td>
</tr>
<tr>
<td>30*</td>
<td>13</td>
<td>3.559E+05</td>
<td>6.647E+04</td>
<td>6.599E+04</td>
<td>3.572E+05</td>
<td>5.196E+04</td>
<td></td>
</tr>
<tr>
<td>31**</td>
<td>10</td>
<td>3.581E+05</td>
<td>6.670E+04</td>
<td>6.618E+04</td>
<td>3.572E+05</td>
<td>5.283E+04</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>9</td>
<td>3.772E+05</td>
<td>6.825E+04</td>
<td>6.516E+04</td>
<td>3.608E+05</td>
<td>4.732E+04</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>6</td>
<td>3.786E+05</td>
<td>6.882E+04</td>
<td>6.660E+04</td>
<td>3.608E+05</td>
<td>4.744E+04</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>1</td>
<td>3.750E+05</td>
<td>6.843E+04</td>
<td>5.388E+04</td>
<td>4.032E+05</td>
<td>4.736E+04</td>
<td></td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 13 terminal nodes
0-SE tree based on median is marked with + and has 28 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on median using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
** tree same as -- tree

Following tree is based on mean CV with naive SE estimate (**)

Structure of final tree. Each terminal node is marked with a T.

D-mean is mean of HIC in the node
Cases fit give the number of cases used to fit node
MSE and R^2 are based on all cases in node

<table>
<thead>
<tr>
<th>Node</th>
<th>Total cases</th>
<th>Cases Matrix</th>
<th>Node fit</th>
<th>D-mean</th>
<th>MSE</th>
<th>R^2</th>
<th>variable</th>
<th>Other variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3276</td>
<td>3272</td>
<td>2</td>
<td>5.127E+02</td>
<td>3.743E+05</td>
<td>0.1007</td>
<td>BARSHP</td>
<td>+YEAR</td>
</tr>
<tr>
<td>2T</td>
<td>40</td>
<td>32</td>
<td>2</td>
<td>1.606E+02</td>
<td>1.337E+04</td>
<td>0.3122</td>
<td>+HW</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3236</td>
<td>3232</td>
<td>2</td>
<td>5.170E+02</td>
<td>3.749E+05</td>
<td>0.1066</td>
<td>BARSHP</td>
<td>+YEAR</td>
</tr>
<tr>
<td>6T</td>
<td>45</td>
<td>45</td>
<td>1</td>
<td>2.603E+02</td>
<td>5.101E+04</td>
<td>0.0000</td>
<td>+Constant*</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>3191</td>
<td>3188</td>
<td>2</td>
<td>5.206E+02</td>
<td>3.757E+05</td>
<td>0.1137</td>
<td>BARSHP</td>
<td>+YEAR</td>
</tr>
<tr>
<td>14</td>
<td>309</td>
<td>204</td>
<td>2</td>
<td>3.486E+02</td>
<td>4.873E+05</td>
<td>0.4317</td>
<td>BX12</td>
<td>-IMPPNT</td>
</tr>
</tbody>
</table>

Wei-Yin Loh

146

GUIDE manual
8.2 Results

PERIODIC VARIABLES: NHTSA DATA

<table>
<thead>
<tr>
<th>Code</th>
<th>Value1</th>
<th>Value2</th>
<th>Value3</th>
<th>Value4</th>
<th>Value5</th>
<th>Value6</th>
</tr>
</thead>
<tbody>
<tr>
<td>28T</td>
<td>44</td>
<td>7</td>
<td>2</td>
<td>7.805E+02</td>
<td>6.238E+05</td>
<td>0.7198 BARDIA</td>
</tr>
<tr>
<td>29T</td>
<td>265</td>
<td>173</td>
<td>2</td>
<td>2.769E+02</td>
<td>2.827E+05</td>
<td>0.5358 SEPOSN IMPPNT</td>
</tr>
<tr>
<td>15</td>
<td>2882</td>
<td>2879</td>
<td>2</td>
<td>5.391E+02</td>
<td>3.315E+05</td>
<td>0.1147 CLSSPD YEAR</td>
</tr>
<tr>
<td>30</td>
<td>1292</td>
<td>9</td>
<td>2</td>
<td>4.440E+02</td>
<td>5.638E+05</td>
<td>0.0267 HS -CB</td>
</tr>
<tr>
<td>60</td>
<td>593</td>
<td>593</td>
<td>1</td>
<td>5.405E+02</td>
<td>5.628E+05</td>
<td>0.0000 RSTFRT <em>Constant</em></td>
</tr>
<tr>
<td>120T</td>
<td>334</td>
<td>82</td>
<td>2</td>
<td>4.267E+02</td>
<td>4.667E+05</td>
<td>0.1066 YEAR IMPPNT</td>
</tr>
<tr>
<td>121T</td>
<td>259</td>
<td>259</td>
<td>1</td>
<td>6.873E+02</td>
<td>5.808E+05</td>
<td>0.0000 VEHCG <em>Constant</em></td>
</tr>
<tr>
<td>61</td>
<td>699</td>
<td>199</td>
<td>2</td>
<td>3.621E+02</td>
<td>5.644E+05</td>
<td>0.0258 CURBWT +CURBWT</td>
</tr>
<tr>
<td>122</td>
<td>591</td>
<td>591</td>
<td>1</td>
<td>3.321E+02</td>
<td>3.861E+05</td>
<td>0.0000 IMPANG <em>Constant</em></td>
</tr>
<tr>
<td>244T</td>
<td>342</td>
<td>341</td>
<td>2</td>
<td>2.158E+02</td>
<td>2.140E+04</td>
<td>0.3209 BX18 +VEHSPD</td>
</tr>
<tr>
<td>245T</td>
<td>249</td>
<td>238</td>
<td>2</td>
<td>4.918E+02</td>
<td>5.705E+05</td>
<td>0.3164 IMPANG +BARDIA</td>
</tr>
<tr>
<td>123T</td>
<td>108</td>
<td>5</td>
<td>2</td>
<td>5.260E+02</td>
<td>1.345E+06</td>
<td>0.1740 VEHLEN -BX3</td>
</tr>
<tr>
<td>31T</td>
<td>1590</td>
<td>1590</td>
<td>2</td>
<td>6.164E+02</td>
<td>1.141E+05</td>
<td>0.4156 MAKED -YEAR</td>
</tr>
</tbody>
</table>

Number of terminal nodes of final tree: 10
Total number of nodes of final tree: 19
Second best split variable (based on curvature test) at root node is IMPANG

Regression tree:
For categorical variable splits, values not in training data go to the right

Node 1: BARSHP = "488", "EOL", "GRL", "MBR", "OTH", "ROR"
  Node 2: HIC-mean = 160.57500
  Node 1: BARSHP /= "488", "EOL", "GRL", "MBR", "OTH", "ROR"
  Node 3: BARSHP = "128", "IAT", "SGN", "UNK"
    Node 6: HIC-mean = 260.28889
    Node 3: BARSHP /= "128", "IAT", "SGN", "UNK"
      Node 7: BARSHP = "134", "EDB", "FAB", "LUM", "US1"
        Node 14: BX12 <= 2552.0000 or BX12 = NA & BX12_ = "A"
          Node 28: HIC-mean = 780.47727
          Node 14: not (BX12 <= 2552.0000 or BX12 = NA & BX12_ = "A")
          Node 29: HIC-mean = 219581579
        Node 15: CLSSPD <= 55.450000
          Node 30: HS <= 325.50000 or NA
            Node 60: RSTFRT = "1"
              Node 120: HIC-mean = 426.69760
              Node 60: RSTFRT /= "1"
              Node 121: HIC-mean = 687.28185
            Node 30: HS > 325.50000
              Node 61: CURBWT <= 1575.0000 or NA
                Node 122: IMPANG in (-1, 1)
                  Node 244: HIC-mean = 215.81579
                  Node 122: IMPANG not in (-1, 1) or NA
                    Node 245: HIC-mean = 491.84337
                    Node 245: HIC-mean = 491.84337
                  Node 61: CURBWT > 1575.0000

Wei-Yin Loh 147 GUIDE manual
8.2 Results

Node 123: HIC-mean = 525.97222
Node 15: CLSSPD > 55.450000 or NA
Node 31: HIC-mean = 616.36541

Predictor means below are means of cases with no missing values. Regression coefficients are computed from the complete cases.

WARNING: p-values below not adjusted for split search. For a bootstrap solution see:


Node 1: Intermediate node
A case goes into Node 2 if BARSHP = "488", "EDL", "GRL", "MBR", "OTH", "ROR"
BARSHP mode = "LCB"

Coefficients of least squares regression function:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.3893E+05</td>
<td>19.40</td>
<td>0.1554E-14</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If regressors have missing values, predicted value = 471.00000
Predicted values truncated at 0.00000 & 12246.0

Node 2: Terminal node

Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-84.23</td>
<td>-0.7127</td>
<td>0.4815</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HW</td>
<td>0.5093</td>
<td>2.369</td>
<td>0.2447E-01</td>
<td>414.0</td>
<td>540.9</td>
<td>787.0</td>
</tr>
</tbody>
</table>

If regressors have missing values, predicted value = 37.875000
Predicted values truncated at 0.00000 & 12246.0

Node 3: Intermediate node
A case goes into Node 6 if BARSHP = "128", "IAT", "SGN", "UNK"
BARSHP mode = "LCB"

Node 6: Terminal node

Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>260.3</td>
<td>7.731</td>
<td>0.9860E-09</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Predicted mean = 260.28889
8.2 Results

Predicted values truncated at 0.00000 & 12246.0

-----------------------------
Node 7: Intermediate node
A case goes into Node 14 if BARSHP = "134", "EOB", "FAB", "LUM", "US1"
BARSHP mode = "LCB"

-----------------------------
Node 14: Intermediate node
A case goes into Node 28 if BX12 <= 2552.0000 or BX12_ = "A"
BX12 mean = 3274.8636

-----------------------------
Node 28: Terminal node
Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.5577E+06</td>
<td>3.041</td>
<td>0.2873E-01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARDIA</td>
<td>-2557.</td>
<td>-3.027</td>
<td>0.2917E-01</td>
<td>216.0</td>
<td>217.1</td>
<td>218.0</td>
</tr>
</tbody>
</table>

If regressors have missing values, predicted value = 469.67568
Predicted values truncated at 0.00000 & 12246.0

-----------------------------
Node 29: Terminal node
Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>146.9</td>
<td>3.108</td>
<td>0.2205E-02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMPPNT</td>
<td>-20.71</td>
<td>-15.12</td>
<td>0.000</td>
<td>-203.0</td>
<td>-5.867</td>
<td>0.000</td>
</tr>
</tbody>
</table>

If regressors have missing values, predicted value = 292.79348
Predicted values truncated at 0.00000 & 12246.0

-----------------------------
Node 15: Intermediate node
A case goes into Node 30 if CLSSPD <= 55.450000
CLSSPD mean = 49.475944

-----------------------------
Node 30: Intermediate node
A case goes into Node 60 if HS <= 325.50000 or NA
HS mean = 323.46129

-----------------------------
Node 60: Intermediate node
A case goes into Node 120 if RSTFRT = "1"
RSTFRT mode = "1"

-----------------------------
Node 120: Terminal node
Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>439.4</td>
<td>2.708</td>
<td>0.8277E-02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMPPNT</td>
<td>-3.173</td>
<td>-2.797</td>
<td>0.6450E-02</td>
<td>-671.0</td>
<td>-62.13</td>
<td>34.00</td>
</tr>
</tbody>
</table>

If regressors have missing values, predicted value = 358.40476
Predicted values truncated at 0.00000 & 12246.0

---

Wei-Yin Loh  149  GUIDE manual
8.2 Results

Node 121: Terminal node
Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>687.3</td>
<td>14.51</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Predicted mean = 687.28185
Predicted values truncated at 0.00000 & 12246.0

Node 61: Intermediate node
A case goes into Node 122 if CURBWT <= 1575.0000 or NA
CURBWT mean = 1689.4824

Node 122: Intermediate node
A case goes into Node 244 if IMPANG in [-1, 1]
IMPANG mean = 114.67005

Node 244: Terminal node
Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-704.0</td>
<td>-9.590</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VEHSPD</td>
<td>21.13</td>
<td>12.61</td>
<td>0.000</td>
<td>23.00</td>
<td>43.55</td>
<td>55.33</td>
</tr>
</tbody>
</table>

If regressors have missing values, predicted value = 57.000000
Predicted values truncated at 0.00000 & 12246.0

Node 245: Terminal node
Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.1118E+07</td>
<td>-11.04</td>
<td>0.3331E-15</td>
<td></td>
<td>254.0</td>
<td>255.0</td>
</tr>
<tr>
<td>BARDIA</td>
<td>4401.</td>
<td>11.04</td>
<td>0.6661E-15</td>
<td>254.0</td>
<td>254.0</td>
<td>255.0</td>
</tr>
</tbody>
</table>

If regressors have missing values, predicted value = 1317.1818
Predicted values truncated at 0.00000 & 12246.0

Node 123: Terminal node
Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.1481E+05</td>
<td>10.91</td>
<td>0.1647E-02</td>
<td></td>
<td>743.0</td>
<td>1210.</td>
</tr>
<tr>
<td>BX3</td>
<td>-12.23</td>
<td>-9.761</td>
<td>0.2284E-02</td>
<td>743.0</td>
<td>1069.</td>
<td>1210.</td>
</tr>
</tbody>
</table>

If regressors have missing values, predicted value = 467.44660
Predicted values truncated at 0.00000 & 12246.0

Node 31: Terminal node
Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.5474E+05</td>
<td>33.99</td>
<td>0.1110E-15</td>
<td></td>
<td>1974.</td>
<td>2017.</td>
</tr>
</tbody>
</table>

If regressors have missing values, predicted value = 616.36541
Predicted values truncated at 0.00000 & 12246.0

Wei-Yin Loh 150 GUIDE manual
Proportion of variance (R-squared) explained by tree model: 0.3546

Observed and fitted values are stored in lin.fit
Regressor names and coefficients are stored in lin.reg
LaTeX code for tree is in lin.tex
R code is stored in lin.r

The piecewise-linear regression tree is shown in Figure 17. The angular split “IMPANG in [-1, 1]” suggests that head-on collision is less serious than otherwise.

9 Poisson regression: solder data

We use a data set on printed circuit board soldering to show how GUIDE fits Poisson regression models. The data were analyzed in Chambers and Hastie (1992) and are given in solder.dat. The description file solder.dsc uses the b descriptor for the 5 categorical variables:

solder.dat
"?"
1
1, skips, d
2, opening, b
3, solder, b
4, mask, b
5, padtype, b
6, panel, b

9.1 Piecewise-constant: solder data

9.1.1 Input file creation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: cons.in
   Input 1 for model fitting, 2 for importance or DIF scoring,
   3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: cons.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1): 2
Figure 17: GUIDE v.41.1 0.250-SE piecewise simple linear least-squares regression tree (constant fitted to incomplete cases in terminal nodes) for predicting HIC. At each split, an observation goes to the left branch if and only if the condition is satisfied. Symbol ‘≤∗’ stands for ‘≤ or missing’. \( S_1 = \{488, \text{EOL, GRL, MBR, OTH, ROR}\} \). \( S_2 = \{128, \text{IAT, SGN, UNK}\} \). \( S_3 = \{134, \text{EOB, FAB, LUM, US1}\} \). Sample size (in italics), mean of HIC, and signed name of regressor variable printed below nodes. Terminal nodes with means above and below value of 512.7 at root node are colored yellow and purple respectively. Asterisk appended to regressor name indicates its slope is significant at the 0.05 level (unadjusted for multiplicity and model search). Second best split variable at root node is IMPANG.
Choose type of regression model:
1=linear, 2=quantile, 3=Poisson, 4=censored response,
5=multiresponse or itemresponse, 6=longitudinal data (with T variables),
7=binary logistic regression.
Input choice ([1:7], <cr>=1): 3
Choose complexity of model to use at each node:
Choose 1 for multiple regression (recommended for prediction)
Choose 2 for best simple polynomial in one N or F variable
Choose 3 for constant fit (recommended for interpretability or if there is an R variable)
1: multiple linear, 2: best simple polynomial, 3: constant ([1:3], <cr>=3):
Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
Input name of data description file (max 100 characters);
enclose with matching quotes if it has spaces: solder.dsc
Reading data description file ... 
Training sample file: solder.dat
Missing value code: NA
Records in data file start on line 1
Warning: B variables changed to C
D variable is skips
Reading data file ...
Number of records in data file: 720
Length of longest entry in data file: 6
Checking for missing values ...
Finished checking
Assigning integer codes to values of 5 categorical variables
Re-checking data ...
Assigning codes to missing values, if any ...
Data checks complete
Number of cases with positive D values: 478
Rereading data ...

<table>
<thead>
<tr>
<th>Total #cases w/</th>
<th>#missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>#cases miss. D</td>
<td>ord. vals</td>
</tr>
<tr>
<td>720</td>
<td>0</td>
</tr>
</tbody>
</table>

No offset variable in data file.
Number of cases used for training: 720
Number of split variables: 5
Finished reading data file
Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):
Input file name to store LaTeX code (use .tex as suffix): cons.tex
You can store the variables and/or values used to split and fit in a file
Choose 1 to skip this step, 2 to store split and fit variables,
3 to store split variables and their values
Input your choice ([1:3], <cr>=1):
Input 2 to save fitted values and node IDs, 1 otherwise ([1:2], <cr>=2):
9.1 Piecewise-constant: solder data

Input name of file to store node ID and fitted value of each case: cons.fit
Input 2 to write R function for predicting new cases, 1 otherwise ([1:2], <cr>=2):
Input file name: cons.r
Input rank of top variable to split root node ([1:5], <cr>=1):
Input file is created!
Run GUIDE with the command: guide < cons.in

The tree is shown in Figure 18, which is quite large. One way to reduce the size of the tree is to fit a more complex Poisson regression model in each node.
Figure 18: GUIDE v.41.1 0.250-SE piecewise constant Poisson regression tree for predicting skips. At each split, an observation goes to the left branch if and only if the condition is satisfied. $S_1 = \{A_{1.5}, A_3\}$. $S_2 = \{D_4, D_7, L_4\}$. $S_3 = \{D_4, D_7, L_4, L_7, L_8\}$. $S_4 = \{L_6, L_9, W_9\}$. $S_5 = \{L_6, L_7, L_9, W_9\}$. $S_6 = \{D_4, D_6, D_7, L_4, W_4\}$. $S_7 = \{D_4, W_4, W_9\}$. $S_8 = \{D_7, L_4, L_8\}$. $S_9 = \{D_4, D_7, L_4, L_8, W_4\}$. $S_{10} = \{D_4, D_7, L_4\}$. Circles with dashed lines are nodes with no significant split variables. Sample size (in italics) and mean of skips printed below nodes. Terminal nodes with means above and below value of 4.97 at root node are colored orange and skyblue respectively. Second best split variable at root node is mask.
9.2 Multiple linear: solder data

Now we construct a tree where each node is fitted with a Poisson model containing only the main effects. This is where the “B” descriptor in solder.dsc is for.

9.2.1 Input file creation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: mul.in
   Input 1 for model fitting, 2 for importance or DIF scoring,
   3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: mul.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1): 2
   Choose type of regression model:
   1=linear, 2=quantile, 3=Poisson, 4=censored response,
   5=multireponse or itemresponse, 6=longitudinal data (with T variables),
   7=binary logistic regression.
   Input choice ([1:7], <cr>=1): 3
   Choose complexity of model to use at each node:
   Choose 1 for multiple regression (recommended for prediction)
   Choose 2 for best simple polynomial in one N or F variable
   Choose 3 for constant fit (recommended for interpretability or if there is an R variable)
   1: multiple linear, 2: best simple polynomial, 3: constant ([1:3], <cr>=3): 1
   Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
   Input name of data description file (max 100 characters); enclose with matching quotes if it has spaces: solder.dsc
   Reading data description file ...
   Training sample file: solder.dat
   Missing value code: NA
   Records in data file start on line 1
   D variable is skips
   Reading data file ...
   Number of records in data file: 720
   Length of longest entry in data file: 6
   Checking for missing values ...
   Finished checking
   Assigning integer codes to values of 5 categorical variables
   Re-checking data ...
   Assigning codes to missing values, if any ...
   Data checks complete
   Number of cases with positive D values: 478
   GUIDE will try to create the variables in the description file.
If it is unsuccessful, please create the columns yourself...
Number of dummy variables created: 17
Creating dummy variables ...
Rereading data ...

<table>
<thead>
<tr>
<th>#cases</th>
<th>miss. D</th>
<th>ord. vals</th>
<th>X-var</th>
<th>N-var</th>
<th>F-var</th>
<th>S-var</th>
<th>P-var</th>
<th>M-var</th>
<th>B-var</th>
<th>C-var</th>
<th>I-var</th>
</tr>
</thead>
<tbody>
<tr>
<td>720</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

No offset variable in data file.
Number of cases used for training: 720
Number of split variables: 5
Number of dummy variables created: 17
Finished reading data file
Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):
Input file name to store LaTeX code (use .tex as suffix): mul.tex
You can store the variables and/or values used to split and fit in a file
Choose 1 to skip this step, 2 to store split and fit variables,
3 to store split variables and their values
Input your choice ([1:3], <cr>=1):
Input 2 to save fitted values and node IDs, 1 otherwise ([1:2], <cr>=2):
Input name of file to store node ID and fitted value of each case: mul.fit
Input 2 to write R function for predicting new cases, 1 otherwise ([1:2], <cr>=2):
Input file name: mul.r
Input rank of top variable to split root node ([1:22], <cr>=1):
Input file is created!
Run GUIDE with the command: guide < mul.in

9.2.2 Contents of mul.out

Poisson regression tree
No truncation of predicted values
Pruning by cross-validation
Data description file: solder.dsc
Training sample file: solder.dat
Missing value code: NA
Records in data file start on line 1
D variable is skips
Piecewise linear model
Number of records in data file: 720
Length of longest entry in data file: 6
Number of cases with positive D values: 478
Number of dummy variables created: 17

Summary information for training sample of size 720
d=dependent, b=split and fit cat var using indicator variables,
9.2 Multiple linear: solder data

\[ c = \text{split-only categorical}, \, i = \text{fit-only categorical (via indicators)}, \]
\[ s = \text{split-only numerical}, \, n = \text{split and fit numerical}, \, f = \text{fit-only numerical}, \]
\[ m = \text{missing-value flag variable}, \, p = \text{periodic variable}, \, w = \text{weight}, \]
\[ z = \text{offset variable} \]

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
<th>#Codes/Levels/</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>skips</td>
<td>d</td>
<td>0.000</td>
<td>48.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>opening</td>
<td>b</td>
<td></td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>solder</td>
<td>b</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>mask</td>
<td>b</td>
<td></td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>padtype</td>
<td>b</td>
<td></td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>panel</td>
<td>b</td>
<td></td>
<td>3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Constructed variables**

| 7      | opening.medium | f       | 0.000   | 1.000   |          |               |
| 8      | opening.small  | f       | 0.000   | 1.000   |          |               |
| 9      | solder.thin    | f       | 0.000   | 1.000   |          |               |
| 10     | mask.A3        | f       | 0.000   | 1.000   |          |               |
| 11     | mask.B3        | f       | 0.000   | 1.000   |          |               |
| 12     | mask.B6        | f       | 0.000   | 1.000   |          |               |
| 13     | padtype.D6     | f       | 0.000   | 1.000   |          |               |
| 14     | padtype.D7     | f       | 0.000   | 1.000   |          |               |
| 15     | padtype.L4     | f       | 0.000   | 1.000   |          |               |
| 16     | padtype.L6     | f       | 0.000   | 1.000   |          |               |
| 17     | padtype.L7     | f       | 0.000   | 1.000   |          |               |
| 18     | padtype.L8     | f       | 0.000   | 1.000   |          |               |
| 19     | padtype.L9     | f       | 0.000   | 1.000   |          |               |
| 20     | padtype.W4     | f       | 0.000   | 1.000   |          |               |
| 21     | padtype.W9     | f       | 0.000   | 1.000   |          |               |
| 22     | panel.2        | f       | 0.000   | 1.000   |          |               |
| 23     | panel.3        | f       | 0.000   | 1.000   |          |               |

Total #cases w/ #missing

<table>
<thead>
<tr>
<th>#cases</th>
<th>miss. D</th>
<th>ord. vals</th>
<th>X-var</th>
<th>N-var</th>
<th>F-var</th>
<th>S-var</th>
<th>P-var</th>
<th>M-var</th>
<th>B-var</th>
<th>C-var</th>
<th>I-var</th>
</tr>
</thead>
<tbody>
<tr>
<td>720</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

No offset variable in data file.

Number of cases used for training: 720

Number of split variables: 5

Number of dummy variables created: 17

Missing regressors imputed with means and missing-value indicators added

Pruning by v-fold cross-validation, with v = 10

Selected tree is based on mean of CV estimates

Number of SE’s for pruned tree: 0.2500

Wei-Yin Loh

GUIDE manual
9.2 Multiple linear: solder data

Nodewise interaction tests on all variables
Fraction of cases used for splitting each node: 1.0000
Maximum number of split levels: 10
Minimum node sample size: 7
Top-ranked variables and chi-squared values at root node
1 0.1782E+02 solder
2 0.3481E+01 opening
3 0.3357E+01 mask
4 0.2453E+00 panel
5 0.1361E+00 padtype

Size and CV Loss and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Nodes</th>
<th>Mean Loss</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median Loss</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>55</td>
<td>2.939E+00</td>
<td>1.916E-01</td>
<td>1.950E-01</td>
<td>2.852E+00</td>
<td>2.525E-01</td>
</tr>
<tr>
<td>2</td>
<td>53</td>
<td>2.939E+00</td>
<td>1.916E-01</td>
<td>1.950E-01</td>
<td>2.852E+00</td>
<td>2.525E-01</td>
</tr>
<tr>
<td>36</td>
<td>4</td>
<td>1.488E+00</td>
<td>8.070E-02</td>
<td>8.672E-02</td>
<td>1.449E+00</td>
<td>7.036E-02</td>
</tr>
<tr>
<td>37**</td>
<td>3</td>
<td>1.457E+00</td>
<td>7.447E-02</td>
<td>9.380E-02</td>
<td>1.343E+00</td>
<td>7.680E-02</td>
</tr>
<tr>
<td>38</td>
<td>2</td>
<td>1.527E+00</td>
<td>7.949E-02</td>
<td>9.597E-02</td>
<td>1.455E+00</td>
<td>6.790E-02</td>
</tr>
<tr>
<td>39</td>
<td>1</td>
<td>1.660E+00</td>
<td>8.239E-02</td>
<td>7.060E-02</td>
<td>1.651E+00</td>
<td>7.689E-02</td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 3 terminal nodes
0-SE tree based on median is marked with + and has 3 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on median and bootstrap SE is marked with ++
* tree, ** tree, + tree, and ++ tree all the same

Following tree is based on mean CV with naive SE estimate (**)

Structure of final tree. Each terminal node is marked with a T.

D-mean is mean of skips in the node
Cases fit give the number of cases used to fit node
Deviance is mean residual deviance for all cases in node

<table>
<thead>
<tr>
<th>Node</th>
<th>Total Cases</th>
<th>Matrix Node</th>
<th>Node</th>
<th>Node Split</th>
<th>Other variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>720</td>
<td>720</td>
<td>18</td>
<td>4.965E+00</td>
<td>1.610E+00 solder</td>
</tr>
<tr>
<td>2T</td>
<td>360</td>
<td>360</td>
<td>17</td>
<td>2.481E+00</td>
<td>1.279E+00 mask</td>
</tr>
<tr>
<td>3</td>
<td>360</td>
<td>360</td>
<td>17</td>
<td>7.450E+00</td>
<td>1.628E+00 opening :mask</td>
</tr>
<tr>
<td>6T</td>
<td>120</td>
<td>120</td>
<td>15</td>
<td>1.636E+01</td>
<td>1.367E+00 padtype</td>
</tr>
<tr>
<td>7T</td>
<td>240</td>
<td>240</td>
<td>16</td>
<td>2.996E+00</td>
<td>1.403E+00 mask</td>
</tr>
</tbody>
</table>

Number of terminal nodes of final tree: 3
Total number of nodes of final tree: 5
Multiple linear: solder data

Second best split variable (based on curvature test) at root node is opening

Regression tree:
For categorical variable splits, values not in training data go to the right

Node 1: solder = "thick"
  Node 2: skips sample mean = 2.4805556
Node 1: solder /= "thick"
  Node 3: opening = "small"
    Node 6: skips sample mean = 16.358333
  Node 3: opening /= "small"
    Node 7: skips sample mean = 2.9958333

******************************************************

WARNING: p-values below not adjusted for split search. For a bootstrap solution see:

Node 1: Intermediate node
A case goes into Node 2 if solder = "thick"
Solder mode = "thick"

Coefficients of regression function for log mean:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-1.220</td>
<td>-12.81</td>
<td>8.882E-15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mask.A3</td>
<td>0.4282</td>
<td>5.674</td>
<td>2.043E-07</td>
<td>0.000</td>
<td>0.250</td>
<td>1.000</td>
</tr>
<tr>
<td>mask.B3</td>
<td>1.202</td>
<td>17.95</td>
<td>7.772E-15</td>
<td>0.000</td>
<td>0.250</td>
<td>1.000</td>
</tr>
<tr>
<td>mask.B6</td>
<td>1.866</td>
<td>29.58</td>
<td>0.000</td>
<td>0.000</td>
<td>0.250</td>
<td>1.000</td>
</tr>
<tr>
<td>opening.medium</td>
<td>0.2585</td>
<td>3.884</td>
<td>1.126E-03</td>
<td>0.000</td>
<td>0.333</td>
<td>1.000</td>
</tr>
<tr>
<td>opening.small</td>
<td>1.893</td>
<td>35.31</td>
<td>8.882E-15</td>
<td>0.000</td>
<td>0.333</td>
<td>1.000</td>
</tr>
<tr>
<td>padtype.D6</td>
<td>-0.3687</td>
<td>-5.164</td>
<td>3.144E-06</td>
<td>0.000</td>
<td>0.100</td>
<td>1.000</td>
</tr>
<tr>
<td>padtype.D7</td>
<td>-9.844E-01</td>
<td>-1.487</td>
<td>0.1374</td>
<td>0.000</td>
<td>0.100</td>
<td>1.000</td>
</tr>
<tr>
<td>padtype.L4</td>
<td>0.2624</td>
<td>4.321</td>
<td>1.774E-04</td>
<td>0.000</td>
<td>0.100</td>
<td>1.000</td>
</tr>
<tr>
<td>padtype.L6</td>
<td>-0.6685</td>
<td>-8.525</td>
<td>0.000</td>
<td>0.000</td>
<td>0.100</td>
<td>1.000</td>
</tr>
<tr>
<td>padtype.L7</td>
<td>-0.4902</td>
<td>-6.619</td>
<td>7.177E-10</td>
<td>0.000</td>
<td>0.100</td>
<td>1.000</td>
</tr>
<tr>
<td>padtype.L8</td>
<td>-0.2712</td>
<td>-3.907</td>
<td>1.023E-03</td>
<td>0.000</td>
<td>0.100</td>
<td>1.000</td>
</tr>
<tr>
<td>padtype.L9</td>
<td>-0.6365</td>
<td>-8.203</td>
<td>2.220E-15</td>
<td>0.000</td>
<td>0.100</td>
<td>1.000</td>
</tr>
<tr>
<td>padtype.W4</td>
<td>-0.1100</td>
<td>-1.657</td>
<td>9.804E-01</td>
<td>0.000</td>
<td>0.100</td>
<td>1.000</td>
</tr>
</tbody>
</table>
### 9.2 Multiple linear: solder data

**Node 2: Terminal node**

Coefficients of regression function for log mean:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-2.431</td>
<td>-10.68</td>
<td>0.000</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>mask.A3</td>
<td>0.4670</td>
<td>2.373</td>
<td>0.1202E-02</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>mask.B3</td>
<td>1.831</td>
<td>11.01</td>
<td>0.000</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>mask.B6</td>
<td>2.520</td>
<td>15.71</td>
<td>0.000</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>opening.medium</td>
<td>0.8641</td>
<td>5.567</td>
<td>0.000</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>opening.small</td>
<td>2.465</td>
<td>18.18</td>
<td>0.000</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>padtype.D6</td>
<td>-0.3238</td>
<td>-2.034</td>
<td>0.1448E-01</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>padtype.D7</td>
<td>0.1201</td>
<td>0.8480</td>
<td>0.3970</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>padtype.L4</td>
<td>0.6985</td>
<td>5.534</td>
<td>0.6221E-07</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>padtype.L6</td>
<td>-0.4002</td>
<td>-2.458</td>
<td>0.1448E-01</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>padtype.L7</td>
<td>0.4167E-01</td>
<td>0.2887</td>
<td>0.7730</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>padtype.L8</td>
<td>0.1481</td>
<td>1.052</td>
<td>0.2936</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>padtype.L9</td>
<td>-0.5921</td>
<td>-3.426</td>
<td>0.6877E-03</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>padtype.W4</td>
<td>-0.5466E-01</td>
<td>-0.3696</td>
<td>0.7119</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>padtype.W9</td>
<td>-1.324</td>
<td>-5.886</td>
<td>0.9394E-08</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>panel.2</td>
<td>0.2224</td>
<td>2.718</td>
<td>0.6895E-02</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>panel.3</td>
<td>0.6825E-01</td>
<td>0.8049</td>
<td>0.4214</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>solder.thin</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
</tbody>
</table>

---

**Node 3: Intermediate node**

A case goes into Node 6 if opening = "small"
opening mode = "large"

---

**Node 6: Terminal node**

Coefficients of regression function for log mean:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>2.080</td>
<td>21.50</td>
<td>0.000</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>mask.A3</td>
<td>0.3085</td>
<td>3.329</td>
<td>0.1202E-02</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>mask.B3</td>
<td>1.050</td>
<td>12.84</td>
<td>0.000</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>mask.B6</td>
<td>1.504</td>
<td>19.34</td>
<td>0.000</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>opening.medium</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>opening.small</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>padtype.D6</td>
<td>-0.2534</td>
<td>-2.788</td>
<td>0.6302E-02</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>padtype.D7</td>
<td>-0.1476</td>
<td>-1.671</td>
<td>0.9763E-01</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>padtype.L4</td>
<td>0.8309E-01</td>
<td>0.9980</td>
<td>0.3206</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>padtype.L6</td>
<td>-0.7187</td>
<td>-6.847</td>
<td>0.4730E-09</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>padtype.L7</td>
<td>-0.6473</td>
<td>-6.315</td>
<td>0.6560E-08</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
<tr>
<td>padtype.L8</td>
<td>-0.4255</td>
<td>-4.452</td>
<td>0.2127E-04</td>
<td>0.000</td>
<td>0.2500</td>
<td>1.000</td>
</tr>
</tbody>
</table>
### 9.3 Offset variable: lung cancer data

We use a data set from an epidemiological study of the effect of public drinking water on cancer mortality in Missouri (Choi et al., 2005). The data file `lungcancer.txt` gives the number of deaths (deaths) from lung cancer among 115 counties (county) during the period 1972–1981 for both sexes (sex) and four age groups (agegp): 45–54, 55–64, 65–74, and over 75. The description file `lungcancer.dsc` below lists...
Our goal is to construct a Poisson regression tree for the gender-specific rate of lung cancer deaths, where rate is the expected number of deaths in a county divided by its population size for each gender. That is, letting \( \mu \) denote the expected number of gender-specific deaths in a county, we fit this model in each node of the tree:

\[
\log(\mu/\text{pop}) = \beta_0 + \beta_1 I(\text{sex} = M).
\]

The variables together with the county population (\( \text{pop} \)) and the natural log of \( \text{pop} \) (\( \text{logpop} \)). The latter is specified as \( z \) to serve as an offset variable and \( \text{pop} \) is excluded (\( x \)) from the analysis. The contents of \texttt{lungcancer.dsc} are:

\begin{verbatim}
lungcancer.txt
NA
1 1 county c
2 sex b
3 agegp c
4 deaths d
5 pop x
6 logpop z
\end{verbatim}

Figure 19: GUIDE v.41.1 0.250-SE multiple linear Poisson regression tree (missing regressor values imputed and missing indicators added) for predicting \textit{skips}. At each split, an observation goes to the left branch if and only if the condition is satisfied. Intermediate nodes with splits due to interaction are in wheat color. Sample size (in italics) and mean of \textit{skips} printed below nodes. Terminal nodes with means above and below value of 4.97 at root node are colored orange and skyblue respectively. Second best split variable at root node is \textit{opening}.
This is achieved by fitting a linear Poisson regression model with sex as \( b \) so that its dummy indicator variable serves as a linear predictor in the Poisson node models.

### 9.3.1 Input file creation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: poi.in
   Input 1 for model fitting, 2 for importance or DIF scoring,
   3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: poi.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1): 2
   Choose type of regression model:
   1=linear, 2=quantile, 3=Poisson, 4=censored response,
   5=multipath or itemresponse, 6=longitudinal data (with T variables),
   7=binary logistic regression.
   Input choice ([1:7], <cr>=1): 3
   Choose complexity of model to use at each node:
   Choose 1 for multiple regression (recommended for prediction)
   Choose 2 for best simple polynomial in one N or F variable
   Choose 3 for constant fit (recommended for interpretability or if there is an R variable)
   1: multiple linear, 2: best simple polynomial, 3: constant ([1:3], <cr>=3): 1
   Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
   Input name of data description file (max 100 characters);
   enclose with matching quotes if it has spaces: lungcancer.dsc
   Reading data description file ...
   Training sample file: lungcancer.txt
   Missing value code: NA
   Records in data file start on line 1
   D variable is deaths
   Reading data file ...
   Number of records in data file: 920
   Length of longest entry in data file: 8
   Checking for missing values ...
   Finished checking
   Assigning integer codes to values of 3 categorical variables
   Re-checking data ...
   Assigning codes to missing values, if any ...
   Data checks complete
   Number of cases with positive D values: 869
   GUIDE will try to create the variables in the description file.
   If it is unsuccessful, please create the columns yourself...
9.3 Offset variable: lung cancer data

Number of dummy variables created: 1
Creating dummy variables ...
Rereading data ...

<table>
<thead>
<tr>
<th>Total #cases w/</th>
<th>#missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>#cases miss. D</td>
<td>0 0 1 0 0 0 0</td>
</tr>
<tr>
<td>#P-var #M-var</td>
<td>0 0 1 2 0</td>
</tr>
</tbody>
</table>

Offset variable in column: 6
Number of cases used for training: 920
Number of split variables: 3
Number of dummy variables created: 1
Finished reading data file
Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):
Input file name to store LaTeX code (use .tex as suffix): poi.tex
You can store the variables and/or values used to split and fit in a file
Choose 1 to skip this step, 2 to store split and fit variables, 3 to store split variables and their values
Input your choice ([1:3], <cr>=1):
Input 2 to save fitted values and node IDs, 1 otherwise ([1:2], <cr>=2):
Input name of file to store node ID and fitted value of each case: poi.fit
Input 2 to write R function for predicting new cases, 1 otherwise ([1:2], <cr>=2):
Input file name: poi.r
Input rank of top variable to split root node ([1:4], <cr>=1):
Input file is created!
Run GUIDE with the command: guide < poi.in

9.3.2 Results

Poisson regression tree
No truncation of predicted values
Pruning by cross-validation
Data description file: lungcancer.dsc
Training sample file: lungcancer.txt
Missing value code: NA
Records in data file start on line 1
D variable is deaths
Piecewise linear model
Number of records in data file: 920
Length of longest entry in data file: 8
Number of cases with positive D values: 869
Number of dummy variables created: 1

Summary information for training sample of size 920
d=dependent, b=split and fit cat variable using indicator variables,
9.3 Offset variable: lung cancer data

\[\text{c=split-only categorical, i=fit-only categorical (via indicators),} \]
\[\text{s=split-only numerical, n=split and fit numerical, f=fit-only numerical,} \]
\[\text{m=missing-value flag variable, p=periodic variable, w=weight,} \]
\[\text{z=offset variable} \]

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 county c</td>
<td></td>
<td>115</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 sex b</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 agegp c</td>
<td></td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 deaths d</td>
<td>0.000</td>
<td>1046.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 logpop z</td>
<td>4.828</td>
<td>10.96</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total #cases w/ #missing

Offset variable in column 6
Number of cases used for training: 920
Number of split variables: 3
Number of dummy variables created: 1

Missing regressors imputed with means and missing-value indicators added
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE's for pruned tree: 0.2500

Node-wise interaction tests on all variables
Fraction of cases used for splitting each node: 1.0000
Maximum number of split levels: 10
Minimum node sample size: 7
Top-ranked variables and 1-df chi-squared values at root node

<table>
<thead>
<tr>
<th>Variable</th>
<th>Chi-Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 agegp</td>
<td>0.2986E+03</td>
</tr>
<tr>
<td>2 sex</td>
<td>0.1574E+02</td>
</tr>
<tr>
<td>3 county</td>
<td>0.7551E-02</td>
</tr>
</tbody>
</table>

Size and CV Loss and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Tnodes</th>
<th>Mean Loss</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median Loss</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>48</td>
<td>2.863E+00</td>
<td>2.930E-01</td>
<td>2.063E-01</td>
<td>2.836E+00</td>
<td>2.864E-01</td>
</tr>
<tr>
<td>2</td>
<td>47</td>
<td>2.863E+00</td>
<td>2.930E-01</td>
<td>2.063E-01</td>
<td>2.836E+00</td>
<td>2.864E-01</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>34</td>
<td>4</td>
<td>2.352E+00</td>
<td>3.233E-01</td>
<td>2.640E-01</td>
<td>2.262E+00</td>
<td>3.370E-01</td>
</tr>
<tr>
<td>35**</td>
<td>3</td>
<td>2.249E+00</td>
<td>3.278E-01</td>
<td>2.705E-01</td>
<td>1.954E+00</td>
<td>2.648E-01</td>
</tr>
</tbody>
</table>

Wei-Yin Loh
36  2  4.702E+00  8.054E-01  4.866E-01  4.153E+00  6.629E-01
37  1  9.431E+00  1.420E+00  9.674E-01  9.043E+00  9.329E-01

0-SE tree based on mean is marked with * and has 3 terminal nodes
0-SE tree based on median is marked with + and has 3 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
* tree, ** tree, + tree, and ++ tree all the same

Following tree is based on mean CV with naive SE estimate (**)

Structure of final tree. Each terminal node is marked with a T.

Rate is mean of Y/exp(offset)
Cases fit give the number of cases used to fit node
Deviance is mean residual deviance for all cases in node

<table>
<thead>
<tr>
<th>Node</th>
<th>Total Cases</th>
<th>Matrix Node</th>
<th>Node</th>
<th>Node Split</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>label</td>
<td>cases</td>
<td>fit</td>
<td>rank</td>
<td>rate</td>
<td>deviance</td>
</tr>
<tr>
<td>1</td>
<td>920</td>
<td>920</td>
<td>2</td>
<td>1.382E-02</td>
<td>9.179E+00</td>
</tr>
<tr>
<td>2T</td>
<td>230</td>
<td>230</td>
<td>2</td>
<td>5.493E-03</td>
<td>1.863E+00</td>
</tr>
<tr>
<td>3</td>
<td>690</td>
<td>690</td>
<td>2</td>
<td>1.763E-02</td>
<td>4.357E+00</td>
</tr>
<tr>
<td>6T</td>
<td>230</td>
<td>230</td>
<td>2</td>
<td>1.339E-02</td>
<td>3.003E+00</td>
</tr>
<tr>
<td>7T</td>
<td>460</td>
<td>460</td>
<td>2</td>
<td>2.093E-02</td>
<td>1.802E+00</td>
</tr>
</tbody>
</table>

Number of terminal nodes of final tree: 3
Total number of nodes of final tree: 5
Second best split variable (based on curvature test) at root node is sex

Regression tree:
For categorical variable splits, values not in training data go to the right

Node 1: agegp = "45-54"
Node 2: deaths sample rate = 0.54928582E-2
Node 1: agegp /= "45-54"
Node 3: agegp = "55-64"
Node 6: deaths sample rate = 0.13389777E-1
Node 3: agegp /= "55-64"
Node 7: deaths sample rate = 0.20932715E-1

******************************************************************************

WARNING: p-values below not adjusted for split search. For a bootstrap solution see:


Wei-Yin Loh 167

GUIDE manual


Node 1: Intermediate node
A case goes into Node 2 if agegp = "45-54"
agegp mode = "45-54"

Coefficients of regression function for log expected rate:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-5.172</td>
<td>-366.9</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>sex.M</td>
<td>1.437</td>
<td>89.64</td>
<td>0.000</td>
<td>0.000</td>
<td>0.5000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Node mean for offset variable = 6.727

----------------------------

Node 2: Terminal node
Coefficients of regression function for log expected rate:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-5.834</td>
<td>-161.5</td>
<td>0.3331E-15</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>sex.M</td>
<td>1.038</td>
<td>24.44</td>
<td>0.2220E-15</td>
<td>0.000</td>
<td>0.5000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Node mean for offset variable = 6.857

----------------------------

Node 3: Intermediate node
A case goes into Node 6 if agegp = "55-64"
agegp mode = "55-64"

----------------------------

Node 6: Terminal node
Coefficients of regression function for log expected rate:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-5.117</td>
<td>-199.8</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>sex.M</td>
<td>1.285</td>
<td>43.87</td>
<td>0.000</td>
<td>0.000</td>
<td>0.5000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Node mean for offset variable = 6.920

----------------------------

Node 7: Terminal node
Coefficients of regression function for log expected rate:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-4.907</td>
<td>-256.9</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>sex.M</td>
<td>1.714</td>
<td>79.68</td>
<td>0.2220E-15</td>
<td>0.000</td>
<td>0.5000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Node mean for offset variable = 6.567

----------------------------

Observed and fitted values are stored in poi.fit
LaTeX code for tree is in poi.tex
R code is stored in poi.r
Figure 20: GUIDE v.41.1 0.250-SE multiple linear Poisson regression tree (constant fitted to incomplete cases in terminal nodes) for predicting rate of deaths. At each split, an observation goes to the left branch if and only if the condition is satisfied. Sample size (in italics) and sample rate printed below nodes. Terminal nodes with rates above and below value of 0.014 at root node are colored orange and skyblue respectively. Second best split variable at root node is sex.

The results show that the death rate increases with age and that the rate for males is consistently higher than that for females. The tree diagram is given in Figure 20.

10 Censored response: RHC data

Section 4 saw the modeling of right heart catheterization (RHC) in terms of the other variables. The data include a time-to-death variable survtime and a variable death that equals 1 if the subject died (uncensored) and equals 0 otherwise (censored). GUIDE can fit a proportional hazards model to the censored survival time if the event indicator death is specified as “D” and survtime as “T”. The description file is rhcdsc2.txt whose contents follow.

```
rhcdsc2.txt
NA
2
1 X x
2 cat1 c
3 cat2 c
4 ca c
5 sadmdte x
```
6 dschdte  
7 dthdte  
8 lstctdte  
9 death  
10 cardiohx  
11 chfhx  
12 dementhx  
13 psychhx  
14 chrpulhx  
15 renalhx  
16 liverhx  
17 gibledhx  
18 malighx  
19 immunhx  
20 transhx  
21 amihx  
22 age  
23 sex  
24 edu  
25 surv2md1  
26 das2d3pc  
27 t3d30  
28 dth30  
29 aps1  
30 scoma1  
31 meanbp1  
32 wbcl1  
33 hrt1  
34 resp1  
35 templ  
36 pafi1  
37 alb1  
38 hema1  
39 bili1  
40 creal  
41 sod1  
42 pot1  
43 paco2l  
44 ph1  
45 swang1  
46 wtkilo1  
47 dnr1  
48 ninsclas  
49 resp  
50 card  
51 neuro  

Wei-Yin Loh

GUIDE manual
10.1 Proportional hazards

GUIDE has two options for modeling censored response data. The first is a piecewise Cox proportional hazards model.

Let the survival time of a subject be $U$ with probability density $f(u)$ and distribution function $F(u)$. The survival probability function is $S(u) = P(U > u) = 1 - F(u)$ and the hazard rate (instantaneous rate of death) at time $u$ is $\lambda(u) = f(u)/S(u)$. Let $U_i$ and $C_i$ be survival and censoring times of subject $i$. Let $Y_i = \min(U_i, C_i)$ be the observed censored survival time and let $\delta_i = I(U_i < C_i)$ denote the event indicator. The proportional hazards model assumes that $\lambda(u, x) = \lambda_0(u) \exp(\beta'x)$, where $\lambda_0(u)$ is an unknown baseline hazard function. Unlike other regression tree methods for survival data, $\lambda_0(u)$ is the same for all terminal nodes of a GUIDE tree.

10.1.1 Input file generation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: censored.in
   Input 1 for model fitting, 2 for importance or DIF scoring,
   3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: censored.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1): 2
   Choose type of regression model:
   1=linear, 2=quantile, 3=Poisson, 4=censored response,
   5=multireponse or itemresponse, 6=longitudinal data (with T variables),
   7=binary logistic regression.
10.1 Proportional hazards

Input choice ([1:7], <cr>=1): 4
Input 1 for proportional hazards, 2 for restricted mean event time ([1:2], <cr>=1):
Choose complexity of model to use at each node:
Choose 1 for multiple regression (recommended for prediction)
Choose 2 for best simple linear in one N or F variable
Choose 3 for constant fit (recommended for interpretability or if there is an R variable)
1: multiple linear, 2: best simple linear, 3: constant ([1:3], <cr>=3):
Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
Input name of data description file (max 100 characters);
enclose with matching quotes if it has spaces: rhcdsc2.txt
Reading data description file ...
Training sample file: rhcdata.txt
Missing value code: NA
Records in data file start on line 2
23 N variables changed to S
D variable is death
Reading data file ...
Number of records in data file: 5735
Length of longest entry in data file: 19
Checking for missing values ...
Finished checking
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Assigning integer codes to values of 31 categorical variables
Finished assigning codes to 10 categorical variables
Finished assigning codes to 20 categorical variables
Finished assigning codes to 30 categorical variables
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...
Finished processing 5000 of 5735 observations
Data checks complete
Smallest uncensored survtime: 2.0000
Number of complete cases excluding censored T < smallest uncensored T: 5735
Number of cases used to compute baseline hazard: 5735
Number of cases with D=1 and T >= smallest uncensored: 3722
Creating missing value indicators ...
Rereading data ...
Largest uncensored and censored survtime with positive weight: 1943.0000 1351.0000

<table>
<thead>
<tr>
<th>Total</th>
<th>#cases w/</th>
<th>#missing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#cases</td>
<td>miss. D</td>
</tr>
<tr>
<td>5735</td>
<td>0</td>
<td>5157</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Survival time variable in column: 64
10.1 Proportional hazards

Event indicator variable in column: 9
Proportion uncensored among nonmissing T and D variables: .649
Number of cases used for training: 5735
Number of split variables: 54
Number of cases excluded due to 0 W or missing D or T variables: 0
Finished reading data file
Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):
Input file name to store LaTeX code (use .tex as suffix): censored.tex
You can store the variables and/or values used to split and fit in a file
Choose 1 to skip this step, 2 to store split and fit variables,
3 to store split variables and their values
Input your choice ([1:3], <cr>=1):
Input 2 to save fitted values and node IDs, 1 otherwise ([1:2], <cr>=2):
Input file name to store node ID and fitted value of each case: censored.fit
Input 2 to write R function for predicting new cases, 1 otherwise ([1:2], <cr>=2):
Input file name: censored.r
Input rank of top variable to split root node ([1:54], <cr>=1):
Input file is created!
Run GUIDE with the command: guide < censored.in

10.1.2 Output file

Regression tree for censored response
Pruning by cross-validation
Data description file: rhcdsc2.txt
Training sample file: rhcdata.txt
Missing value code: NA
Records in data file start on line 2
23 N variables changed to S
D variable is death
Piecewise constant model
Number of records in data file: 5735
Length of longest entry in data file: 19
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Number of complete cases excluding censored T < smallest uncensored T: 5735
Number of cases used to compute baseline hazard: 5735
Number of cases with D=1 and T >= smallest uncensored: 3722
Smallest uncensored survtime: 2.0000
Largest uncensored and censored survtime with positive weight: 1943.0000 1351.0000

Summary information for training sample of size 5735
d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators),
s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
10.1 Proportional hazards

m=missing-value flag variable, p=periodic variable, w=weight, t=survival time variable

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>cat1</td>
<td>c</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>cat2</td>
<td>c</td>
<td>6</td>
<td>4535</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>ca</td>
<td>c</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>death</td>
<td>d</td>
<td>0.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>ortho</td>
<td>c</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>addd3p</td>
<td>s</td>
<td>0.000</td>
<td>7.000</td>
<td>4296</td>
</tr>
<tr>
<td>60</td>
<td>urin1</td>
<td>s</td>
<td>0.000</td>
<td>9000.</td>
<td>3028</td>
</tr>
<tr>
<td>61</td>
<td>race</td>
<td>c</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>income</td>
<td>c</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>survtime</td>
<td>t</td>
<td>2.000</td>
<td>1943.</td>
<td></td>
</tr>
</tbody>
</table>

=================== Constructed variables ===================

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>lnbasehaz</td>
<td>z</td>
<td>-3.818</td>
</tr>
</tbody>
</table>

Total #cases w/ #missing

<table>
<thead>
<tr>
<th>#cases</th>
<th>miss. D ord. vals</th>
<th>#X-var</th>
<th>#N-var</th>
<th>#F-var</th>
<th>#S-var</th>
</tr>
</thead>
<tbody>
<tr>
<td>5735</td>
<td>0</td>
<td>5157</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#P-var</th>
<th>#M-var</th>
<th>#B-var</th>
<th>#C-var</th>
<th>#I-var</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>31</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Survival time variable in column: 64
Event indicator variable in column: 9
Proportion uncensored among nonmissing T and D variables: 0.649
Number of cases used for training: 5735
Number of split variables: 54
Number of cases excluded due to 0 W or missing D or T variables: 0

Constant fitted to cases with missing values in regressor variables
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE's for pruned tree: 0.2500

Nodewise interaction tests on all variables
Fraction of cases used for splitting each node: 1.0000
Maximum number of split levels: 15
Minimum node sample size: 57
Number of iterations for fitting: 20
Top-ranked variables and 1-df chi-squared values at root node

<table>
<thead>
<tr>
<th>Rank</th>
<th>Chi-Squared Value</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.7573E+03</td>
<td>surv2md1</td>
</tr>
<tr>
<td>2</td>
<td>0.3288E+03</td>
<td>addd3p</td>
</tr>
<tr>
<td>3</td>
<td>0.2341E+03</td>
<td>cat1</td>
</tr>
<tr>
<td>4</td>
<td>0.2263E+03</td>
<td>aps1</td>
</tr>
</tbody>
</table>

Wei-Yin Loh 174 GUIDE manual
10.1 Proportional hazards

CENSORED RESPONSE: RHC DATA

Size and CV Loss and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Tnodes</th>
<th>Mean Loss</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median Loss</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>74</td>
<td>1.284E+00</td>
<td>1.996E-02</td>
<td>1.216E-02</td>
<td>1.282E+00</td>
<td>1.261E-02</td>
</tr>
<tr>
<td>2</td>
<td>73</td>
<td>1.284E+00</td>
<td>1.996E-02</td>
<td>1.228E-02</td>
<td>1.282E+00</td>
<td>1.262E-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>11</td>
<td>1.251E+00</td>
<td>1.800E-02</td>
<td>1.319E-02</td>
<td>1.251E+00</td>
<td>1.993E-02</td>
</tr>
<tr>
<td>44**</td>
<td>10</td>
<td>1.246E+00</td>
<td>1.776E-02</td>
<td>1.259E-02</td>
<td>1.237E+00</td>
<td>1.786E-02</td>
</tr>
<tr>
<td>45++</td>
<td>8</td>
<td>1.254E+00</td>
<td>1.718E-02</td>
<td>1.245E-02</td>
<td>1.241E+00</td>
<td>1.868E-02</td>
</tr>
<tr>
<td>46</td>
<td>7</td>
<td>1.259E+00</td>
<td>1.717E-02</td>
<td>1.177E-02</td>
<td>1.249E+00</td>
<td>2.188E-02</td>
</tr>
<tr>
<td>47</td>
<td>6</td>
<td>1.273E+00</td>
<td>1.723E-02</td>
<td>1.130E-02</td>
<td>1.270E+00</td>
<td>1.882E-02</td>
</tr>
<tr>
<td>48</td>
<td>5</td>
<td>1.289E+00</td>
<td>1.744E-02</td>
<td>1.194E-02</td>
<td>1.284E+00</td>
<td>1.923E-02</td>
</tr>
<tr>
<td>49</td>
<td>3</td>
<td>1.296E+00</td>
<td>1.714E-02</td>
<td>1.295E-02</td>
<td>1.297E+00</td>
<td>2.324E-02</td>
</tr>
<tr>
<td>50</td>
<td>2</td>
<td>1.337E+00</td>
<td>1.699E-02</td>
<td>1.161E-02</td>
<td>1.331E+00</td>
<td>1.397E-02</td>
</tr>
<tr>
<td>51</td>
<td>1</td>
<td>1.459E+00</td>
<td>1.629E-02</td>
<td>6.178E-03</td>
<td>1.454E+00</td>
<td>9.978E-03</td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 10 terminal nodes
Selected-SE tree based on median using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
* tree same as + tree
** tree same as + tree
** tree same as -- tree
* tree same as ** tree
* tree same as -- tree

Following tree is based on mean CV with naive SE estimate (**)
Number of terminal nodes of final tree: 10
Total number of nodes of final tree: 19
Second best split variable (based on curvature test) at root node is adld3p

Regression tree:
For categorical variable splits, values not in training data go to the right

Node 1: surv2md1 <= 0.56447053
  Node 2: adld3p = NA
    Node 4: surv2md1 <= 0.35847378
      Node 8: Median survival time = 11.000000
      Node 4: surv2md1 > 0.35847378 or NA
        Node 9: dnr1 = "No"
          Node 18: Median survival time = 37.000000
          Node 9: dnr1 /= "No"
            Node 19: Median survival time = 8.000000
        Node 2: adld3p /= NA
          Node 5: Median survival time = 195.000000
    Node 2: adld3p /= NA
    Node 3: surv2md1 <= 0.71744752
      Node 6: adld3p = NA
        Node 12: dnr1 = "No"
          Node 24: Median survival time = 171.000000
          Node 12: dnr1 /= "No"
            Node 25: Median survival time = 25.500000
        Node 6: adld3p /= NA
          Node 13: Median survival time = 375.000000
      Node 3: surv2md1 > 0.71744752 or NA
        Node 7: chfhx = "0"
          Node 14: das2d3pc <= 23.857420
            Node 28: Median survival time = 364.000000
            Node 14: das2d3pc > 23.857420 or NA
              Node 29: Median survival time = 672.000000
              Node 7: chfhx /= "0"
10.1 Proportional hazards

Node 15: Median survival time = 373.00000

***********************************************************************
Predictor means below are means of cases with no missing values.

WARNING: p-values below not adjusted for split search. For a bootstrap solution see:


Node 1: Intermediate node
A case goes into Node 2 if surv2md1 <= 0.56447053
surv2md1 mean = 0.59245008
Coefficients of log-relative hazard function (relative to baseline hazard):
Regressor     Coefficient   t-stat   p-value
Constant      0.000
Predicted log-relative hazard = 192.000000

Node 2: Intermediate node
A case goes into Node 4 if adld3p = NA
adld3p mean = 1.3589744

Node 4: Intermediate node
A case goes into Node 8 if surv2md1 <= 0.35847378
surv2md1 mean = 0.38175857

Node 8: Terminal node
Coefficients of log-relative hazard function (relative to baseline hazard):
Regressor     Coefficient   t-stat   p-value
Constant      1.015
Predicted log-relative hazard = 11.000000

Node 15: Terminal node
Coefficients of log-relative hazard function (relative to baseline hazard):
Regressor     Coefficient   t-stat   p-value
Constant     -0.4135
Predicted log-relative hazard = 373.000000

Observed and fitted values are stored in censored.fit
Figure 21: GUIDE v.41.1 0.250-SE piecewise-constant proportional hazards regression tree for survtime. At each split, an observation goes to the left branch if and only if the condition is satisfied. Sample size (in italics), relative hazard, and median survival time printed below nodes. Terminal nodes with median survival times above and below 192 (median at root node) are colored yellow and vermilion respectively. Second best split variable at root node is adld3p.

LaTeX code for tree is in censored.tex
R code is stored in censored.r

The top few lines of the file censored.fit are:

<table>
<thead>
<tr>
<th>train</th>
<th>node</th>
<th>observed</th>
<th>event</th>
<th>logbasecumhaz</th>
<th>survivalprob</th>
<th>median</th>
<th>survtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>13</td>
<td>240.000</td>
<td>n</td>
<td>-0.261185</td>
<td>0.631158</td>
<td>375.000</td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>15</td>
<td>45.0000</td>
<td>y</td>
<td>-0.804384</td>
<td>0.743903</td>
<td>373.000</td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>8</td>
<td>317.000</td>
<td>y</td>
<td>-0.500244E-001</td>
<td>0.725445E-001</td>
<td>11.0000</td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>18</td>
<td>37.0000</td>
<td>y</td>
<td>-0.889004</td>
<td>0.553180</td>
<td>37.0000</td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>19</td>
<td>2.000000</td>
<td>y</td>
<td>-4.01055</td>
<td>0.943144</td>
<td>8.00000</td>
<td></td>
</tr>
</tbody>
</table>

The columns are:

Wei-Yin Loh

178

GUIDE manual
10.1 Proportional hazards

**train**: equals y if observation is used for model fitting; equals n if not used.

**node**: terminal node label of observation.

**observed**: observed survival time (t variable in description file).

**event**: equals y if observed is uncensored (d=1); equals n if censored (d=0).

**logbasecumhaz**: log of the estimated baseline cumulative hazard function \( \log \Lambda_0(t) = \log \int_0^t \lambda_0(u) \, du \) at observed time \( t \).

**survivalprob**: probability that the subject survives up to observed time \( t \). For the first subject, this is

\[
\exp\{-\Lambda_0(t) \exp(\beta'x)\} = \exp\{-\exp(\beta_0 + \text{logbasecumhaz})\} \\
= \exp\{-\exp(-0.514911594896 - 0.261185)\} \\
= 0.6311581
\]

where \( t = 240 \) and \( \beta_0 = -0.514911594896 \) is the constant term in the node (censored.r gives \( \beta_0 \) to higher precision than censored.out).

**mediansurvtime**: median survival time among observations in node estimated from Kaplan-Meier survival function. A trailing plus (+) sign indicates estimate is censored.

Figure 22 plots the estimated survival curves in the terminal nodes of the tree. The plot is produced by the following R code.

```r
library(survival)
z0 <- read.table("rhcdata.txt",header=TRUE)
z1 <- read.table("censored.fit",header=TRUE)
node <- unique(sort(z1$node))
leg.txt <- paste("Node",node)
leg.col <- c("green","magenta","blue","cyan","red")
leg.lty <- rep(c(1,2),c(5,5))
fit <- survfit(Surv(z0$survtime,z0$death) ~ z1$node, conf.type="none")
plot(fit,mark.time=FALSE,xlab="Survival time",ylab="Survival probability", 
    col=leg.col,lwd=2,lty=leg.lty)
title("Kaplan-Meier survival curves")
legend("topright",legend=leg.txt,lty=leg.lty,col=leg.col,lwd=2,ncol=2)
```
Figure 22: Kaplan-Meier survival curves for data in terminal nodes of Figure 21
10.2 Restricted mean event time

The mean survival time is not estimable if there is censoring. But given a pre-specified time point $\tau$, the restricted mean survival time $\mu(X) = E(Y|X)$ is estimable, where $Y = \min(U, C, \tau)$ and $X$ is a covariate vector (Andersen et al., 2004; Chen and Tsiatis, 2001; Tian et al., 2014). GUIDE has an option to fit a restricted event time model to each node of the tree such that $\mu(X)$ is linear in the covariates.

10.2.1 Input file creation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: rest.in
   Input 1 for model fitting, 2 for importance or DIF scoring,
   3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: rest.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1): 2
   Choose type of regression model:
   1=linear, 2=quantile, 3=Poisson, 4=censored response,
   5=multireponse or itemresponse, 6=longitudinal data (with T variables),
   7=binary logistic regression.
   Input choice ([1:7], <cr>=1): 4
   Input 1 for proportional hazards, 2 for restricted mean event time ([1:2], <cr>=1): 2
   Choose complexity of model to use at each node:
   Choose 0 for stepwise linear regression (recommended for prediction)
   Choose 1 for multiple regression
   Choose 2 for best simple polynomial in one N or F variable
   Choose 3 for constant fit (recommended for interpretability or if there is an R variable)
   0: stepwise linear, 1: multiple linear, 2: best simple polynomial, 3: constant,
   4: best simple stepwise ANCOVA ([0:4], <cr>=3):
   Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
   Input name of data description file (max 100 characters);
   enclose with matching quotes if it has spaces: rhcdsc2.txt
   Reading data description file ...
   Training sample file: rhcdata.txt
   Missing value code: NA
   Records in data file start on line 2
   23 N variables changed to S
   D variable is death
   Reading data file ...
   Number of records in data file: 5735
   Length of longest entry in data file: 19

Wei-Yin Loh 181 GUIDE manual
Checking for missing values ...
Finished checking
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Assigning integer codes to values of 31 categorical variables
Finished assigning codes to 10 categorical variables
Finished assigning codes to 20 categorical variables
Finished assigning codes to 30 categorical variables
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...
Finished processing 5000 of 5735 observations
Data checks complete
Creating missing value indicators ...
Rereading data ...
Largest uncensored and censored survtme with positive weight: 1943.0000 1351.0000
Smallest observed uncensored time is 2.0000
Largest observed censored or uncensored time is 1943.0000
Input restriction on event time ([2.00:1943.00], <cr>=972.00):

<table>
<thead>
<tr>
<th>Total</th>
<th>#cases w/</th>
<th>#missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>#cases miss. D ord. vals</td>
<td>X-var</td>
<td>N-var</td>
</tr>
<tr>
<td>5735</td>
<td>0</td>
<td>5157</td>
</tr>
<tr>
<td>#P-var</td>
<td>M-var</td>
<td>B-var</td>
</tr>
<tr>
<td>0 0</td>
<td>0</td>
<td>31</td>
</tr>
</tbody>
</table>

No weight variable in data file
Number of cases used for training: 3732
Number of split variables: 54
Number of cases excluded due to 0 W or missing D variable: 2003
Finished reading data file
Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):
Input file name to store LaTeX code (use .tex as suffix): rest.tex
You can store the variables and/or values used to split and fit in a file
Choose 1 to skip this step, 2 to store split and fit variables,
3 to store split variables and their values
Input your choice ([1:3], <cr>=1):
Input 2 to save fitted values and node IDs, 1 otherwise ([1:2], <cr>=2):
Input name of file to store node ID and fitted value of each case: rest.fit
Input 2 to write R function for predicting new cases, 1 otherwise ([1:2], <cr>=2):
Input file name: rest.r
Input rank of top variable to split root node ([1:54], <cr>=1):
Run GUIDE with the command: guide < rest.in

Wei-Yin Loh 182 GUIDE manual
10.2.2 Contents of rest.out

Restricted mean event time regression tree
Pruning by cross-validation
Data description file: rhcsc2.txt
Training sample file: rhcdata.txt
Missing value code: NA
Records in data file start on line 2
23 N variables changed to S
D variable is death
Piecewise constant model
Number of records in data file: 5735
Length of longest entry in data file: 19
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Smallest uncensored survtime: 2.0000
Largest uncensored and censored survtime with positive weight: 1943.0000 1351.0000
Interval for restricted mean event time is from 0 to 972.

Summary information for training sample of size 3732 (excluding observations with non-positive weight or missing values in d, e, t, r or z variables)

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>cat1</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>cat2</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>ca</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>death</td>
<td>d</td>
<td>0.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>race</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>income</td>
<td>c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>survtime</td>
<td>t</td>
<td>2.000</td>
<td>1943.</td>
<td></td>
</tr>
</tbody>
</table>

Total #cases w/ #missing
#cases miss. D ord. vals #X-var #N-var #F-var #S-var
5735 0 5157 8 0 0 23
#P-var #M-var #B-var #C-var #I-var
0 0 0 31 0

No weight variable in data file
Number of cases used for training: 3732
Number of split variables: 54
10.2 Restricted mean event time

Number of cases excluded due to 0 W or missing D variable: 2003

Constant fitted to cases with missing values in regressor variables
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE’s for pruned tree: 0.2500

Node-wise interaction tests on all variables
Split values for N and S variables based on exhaustive search
Maximum number of split levels: 13
Minimum node sample size: 37
Top-ranked variables and 1-df chi-squared values at root node
  1 0.1868E+03 adld3p
  2 0.1629E+03 surv2md1
  3 0.1122E+03 cat1
  :
  51 0.1196E+00 amihx
  52 0.6209E-01 income

Size and CV MSE and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Nodes</th>
<th>Mean MSE</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median MSE</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>75</td>
<td>1.121E+05</td>
<td>3.376E+03</td>
<td>2.477E+03</td>
<td>1.120E+05</td>
<td>2.101E+03</td>
</tr>
<tr>
<td>2</td>
<td>74</td>
<td>1.121E+05</td>
<td>3.376E+03</td>
<td>2.477E+03</td>
<td>1.119E+05</td>
<td>2.107E+03</td>
</tr>
<tr>
<td>:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>10</td>
<td>1.105E+05</td>
<td>3.343E+03</td>
<td>2.142E+03</td>
<td>1.106E+05</td>
<td>2.773E+03</td>
</tr>
<tr>
<td>43+</td>
<td>8</td>
<td>1.086E+05</td>
<td>3.212E+03</td>
<td>2.008E+03</td>
<td>1.082E+05</td>
<td>3.190E+03</td>
</tr>
<tr>
<td>44</td>
<td>7</td>
<td>1.086E+05</td>
<td>3.184E+03</td>
<td>2.177E+03</td>
<td>1.086E+05</td>
<td>3.279E+03</td>
</tr>
<tr>
<td>45**</td>
<td>6</td>
<td>1.067E+05</td>
<td>3.063E+03</td>
<td>1.467E+03</td>
<td>1.084E+05</td>
<td>2.196E+03</td>
</tr>
<tr>
<td>46</td>
<td>4</td>
<td>1.091E+05</td>
<td>3.044E+03</td>
<td>1.503E+03</td>
<td>1.090E+05</td>
<td>2.580E+03</td>
</tr>
<tr>
<td>47</td>
<td>3</td>
<td>1.097E+05</td>
<td>3.045E+03</td>
<td>1.425E+03</td>
<td>1.090E+05</td>
<td>1.927E+03</td>
</tr>
<tr>
<td>48</td>
<td>2</td>
<td>1.102E+05</td>
<td>3.062E+03</td>
<td>1.527E+03</td>
<td>1.102E+05</td>
<td>2.279E+03</td>
</tr>
<tr>
<td>49</td>
<td>1</td>
<td>1.225E+05</td>
<td>3.100E+03</td>
<td>2.805E+02</td>
<td>1.225E+05</td>
<td>4.687E+02</td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 6 terminal nodes
0-SE tree based on median is marked with + and has 8 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
** tree same as ++ tree
** tree same as -- tree
++ tree same as -- tree
* tree same as ** tree
* tree same as ++ tree
* tree same as -- tree

Following tree is based on mean CV with naive SE estimate (**)
Structure of final tree. Each terminal node is marked with a T.

D-mean is weighted mean of death in the node
Cases fit give the number of cases used to fit node
MSE is residual sum of squares divided by number of cases in node

<table>
<thead>
<tr>
<th>Node</th>
<th>Total Cases</th>
<th>Matrix Cases</th>
<th>Node Split Variable</th>
<th>Interacting Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3732</td>
<td>3732</td>
<td>1</td>
<td>3.144E+02</td>
</tr>
<tr>
<td>2</td>
<td>664</td>
<td>664</td>
<td>1</td>
<td>4.685E+02</td>
</tr>
<tr>
<td>4T</td>
<td>168</td>
<td>168</td>
<td>1</td>
<td>3.244E+02</td>
</tr>
<tr>
<td>5</td>
<td>496</td>
<td>496</td>
<td>1</td>
<td>5.040E+02</td>
</tr>
<tr>
<td>10T</td>
<td>314</td>
<td>314</td>
<td>1</td>
<td>5.756E+02</td>
</tr>
<tr>
<td>11T</td>
<td>182</td>
<td>182</td>
<td>1</td>
<td>3.515E+02</td>
</tr>
<tr>
<td>3</td>
<td>3068</td>
<td>3068</td>
<td>1</td>
<td>2.647E+02</td>
</tr>
<tr>
<td>6T</td>
<td>1262</td>
<td>1262</td>
<td>1</td>
<td>1.607E+02</td>
</tr>
<tr>
<td>7</td>
<td>1806</td>
<td>1806</td>
<td>1</td>
<td>3.225E+02</td>
</tr>
<tr>
<td>14T</td>
<td>1000</td>
<td>1000</td>
<td>1</td>
<td>4.001E+02</td>
</tr>
<tr>
<td>15T</td>
<td>806</td>
<td>806</td>
<td>1</td>
<td>2.057E+02</td>
</tr>
</tbody>
</table>

Number of terminal nodes of final tree: 6
Total number of nodes of final tree: 11
Second best split variable (based on curvature test) at root node is surv2md1

Regression tree:

Node 1: adld3p <= 5.5000000
  Node 2: surv2md1 <= 0.58646870
    Node 4: survtime-mean = 324.40508
    Node 2: surv2md1 > 0.58646870 or NA
    Node 5: urin1 = NA
      Node 10: survtime-mean = 575.62515
      Node 5: urin1 /= NA
        Node 11: survtime-mean = 351.45397
    Node 2: surv2md1 > 0.58646870 or NA
      Node 5: urin1 /= NA
        Node 11: survtime-mean = 351.45397
  Node 1: adld3p > 5.5000000 or NA
    Node 3: surv2md1 <= 0.49098337
      Node 6: survtime-mean = 160.70095
      Node 3: surv2md1 > 0.49098337 or NA
        Node 7: urin1 = NA
          Node 14: survtime-mean = 400.06348
          Node 7: urin1 /= NA
            Node 15: survtime-mean = 205.70770

******************************************************************************
Predictor means below are means of cases with no missing values.
### 10.2 Restricted mean event time

**WARNING:** p-values below not adjusted for split search. For a bootstrap solution see:


**Node 1:** Intermediate node  
A case goes into Node 2 if \( \text{adj3p} \leq 5.500000 \)  
\( \text{adj3p mean} = 1.2733830 \)  
Coefficients of least squares regression function:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>314.4</td>
<td>45.27</td>
<td>0.000</td>
</tr>
<tr>
<td>( \text{survtime} ) mean = 314.380</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Node 2:** Intermediate node  
A case goes into Node 4 if \( \text{surv2md1} \leq 0.58646870 \)  
\( \text{surv2md1 mean} = 0.68493485 \)

**Node 4:** Terminal node  
Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>324.4</td>
<td>11.22</td>
<td>0.000</td>
</tr>
<tr>
<td>( \text{survtime} ) mean = 324.405</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Node 5:** Intermediate node  
A case goes into Node 10 if \( \text{urin1} = \text{NA} \)  
\( \text{urin1 mean} = 2420.9321 \)

**Node 10:** Terminal node  
Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>575.6</td>
<td>19.18</td>
<td>0.000</td>
</tr>
<tr>
<td>( \text{survtime} ) mean = 575.625</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Node 11:** Terminal node  
Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>351.5</td>
<td>14.47</td>
<td>0.000</td>
</tr>
<tr>
<td>( \text{survtime} ) mean = 351.454</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Node 3:** Intermediate node

---

Wei-Yin Loh  
186  
GUIDE manual
A case goes into Node 6 if \( \text{surv2md1} \leq 0.49098337 \)
\( \text{surv2md1 mean} = 0.54259828 \)

Node 6: Terminal node

Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>160.7</td>
<td>19.16</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\( \text{survtime mean} = 160.701 \)

Node 7: Intermediate node

A case goes into Node 14 if \( \text{urin1} = \text{NA} \)
\( \text{urin1 mean} = 1998.7301 \)

Node 14: Terminal node

Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>400.1</td>
<td>25.39</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\( \text{survtime mean} = 400.063 \)

Node 15: Terminal node

Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>205.7</td>
<td>20.34</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\( \text{survtime mean} = 205.708 \)

Observed and fitted values are stored in rest.fit
LaTeX code for tree is in rest.tex
R code is stored in rest.r

Figure 23 shows the restricted mean event time tree.

11 Randomized treatments

Causal effects of treatments are best studied in a randomized trial where the treatments are assigned randomly to subjects. The goal is to show that one treatment is more efficacious than another across all subjects. If this determination is not achieved, a secondary goal may be to search for subgroups of subjects with differential treatment effects.

There are two types of covariates for identification of subgroups with differential treatment effects. A *prognostic* variable is a clinical or biologic characteristic that provides information on the likely outcome of the disease in an untreated individual (e.g., patient age, family history, disease stage, and prior therapy). A *predictive*
Figure 23: GUIDE v.41.1 0.250-SE piecewise-constant regression tree for mean `survtime` restricted to less than 972.000. At each split, an observation goes to the left branch if and only if the condition is satisfied. Sample size (in italics) and restricted mean of `survtime` printed below nodes. Terminal nodes with means above and below value of 314.4 at root node are colored yellow and skyblue respectively. Second best split variable at root node is `surv2md1`. 

Wei-Yin Loh

188

GUIDE manual
variable is one that provides information on the likely benefit from the treatment.
Predictive variables can be used to identify subgroups of patients who are most likely
to benefit from a given therapy. In general, prognostic variables define the effects of
patient or tumor characteristics on the patient outcome, whereas predictive variables
define the effect of treatment on the tumor (Italiano, 2011). Accordingly, GUIDE
has two options, called Gi and Gs. Gi is more sensitive to predictive variables and
Gs tends to be equally sensitive to prognostic and predictive variables (Loh et al.,
2015).

11.1 Multiple treatment arms: CAPE data

We first demonstrate this on a data set from a three-armed randomized controlled
experiment to find out whether two interventions (DVD or Phone) are more efficac-
cious than a control at promoting mammography screening. The relevant data and
description files are cape.txt and cape.dsc. Note that the three treatment levels
(contained in the treatment (R) variable group) are assumed to be categorical (i.e.,
nominal valued). See Loh et al. (2016) for more information on the data.

Because the response variable (resp6) is 0-1 (0=no, 1=yes), we use least-squares
regression with resp6 designated as the dependent variable D or d in the description
file. The treatment variable (group) is designated as R or r (for “Rx”).

11.1.1 Input file creation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: gi.in
   Input 1 for model fitting, 2 for importance or DIF scoring,
   3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: gi.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1): 2
   Choose type of regression model:
   1=linear, 2=quantile, 3=Poisson, 4=censored response,
   5=multiresponse or itemresponse, 6=longitudinal data (with T variables),
   7=binary logistic regression.
   Input choice ([1:7], <cr>=1):
   Input 1 for least squares, 2 least median of squares ([1:2], <cr>=1):
   Choose complexity of model to use at each node:
   Choose 0 for stepwise linear regression (recommended for prediction)
   Choose 1 for multiple regression
11.1 Multiple treatment arms: CAPE data

Choose 2 for best simple polynomial in one N or F variable
Choose 3 for constant fit (recommended for interpretability or if there is an R variable)
0: stepwise linear, 1: multiple linear, 2: best simple polynomial, 3: constant,
4: best simple stepwise ANCOVA ([0:4], <cr>=3):
Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
Input name of data description file (max 100 characters);
enclose with matching quotes if it has spaces: cape.dsc

Reading data description file ...
Training sample file: cape.txt
Missing value code: NA
Records in data file start on line 1
R variable present
21 N variables changed to S
Warning: model changed to linear in treatment
D variable is resp6
Reading data file ...
Number of records in data file: 1681
Length of longest entry in data file: 25
Checking for missing values ...
Finished checking
Missing values found in D variable
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Assigning integer codes to values of 18 categorical variables
Finished assigning codes to 10 categorical variables
Treatment (R) variable is group with values "Control", "DVD", and "Phone"
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...
Data checks complete
GUIDE will try to create the variables in the description file.
If it is unsuccessful, please create the columns yourself...
Number of dummy variables created: 2
Choose a subgroup identification method:
1 = Prognostic priority (Gs)
2 = Predictive priority (Gi)
Input your choice: ([1:2], <cr>=2):
Creating dummy variables ...
Creating missing value indicators ...
Rereading data ...
Proportion of training sample for each level of group
"Control" 0.3278
"DVD" 0.3309
"Phone" 0.3413
Total #cases w/ #missing
### 11.1 Multiple treatment arms: CAPE data

| #cases miss. D ord. vals #X-var #N-var #F-var #S-var | 
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| 1681            | 43              | 84              | 1               | 0               | 0               | 21              |

No weight variable in data file

Number of cases used for training: 1638

Number of split variables: 38

Number of dummy variables created: 2

Number of cases excluded due to 0 W or missing D or R variables: 43

Finished reading data file

Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):

Input file name to store LaTeX code (use .tex as suffix): gi.tex

You can store the variables and/or values used to split and fit in a file

Choose 1 to skip this step, 2 to store split and fit variables, 3 to store split variables and their values

Input your choice ([1:3], <cr>=1):

Input 2 to save fitted values and node IDs, 1 otherwise ([1:2], <cr>=2):

Input name of file to store node ID and fitted value of each case: gi.fit

Input 2 to write R function for predicting new cases, 1 otherwise ([1:2], <cr>=2):

Input file name: gi.r

Input rank of top variable to split root node ([1:41], <cr>=1):

Input file is created!

Run GUIDE with the command: guide < gi.in

#### 11.1.2 Contents of gi.out

Least squares regression tree

Pruning by cross-validation

Data description file: cape.dsc

Training sample file: cape.txt

Missing value code: NA

Records in data file start on line 1

R variable present

21 N variables changed to S

Warning: model changed to linear in treatment

D variable is resp6

Piecewise linear model

Number of records in data file: 1681

Length of longest entry in data file: 25

Missing values found in D variable

Missing values found among categorical variables

Separate categories will be created for missing categorical variables

Missing values found among non-categorical variables

Treatment (R) variable is group with values "Control", "DVD", and "Phone"

Number of dummy variables created: 2

Proportion of training sample for each level of group
11.1 Multiple treatment arms: CAPE data

"Control" 0.3278
"DVD" 0.3309
"Phone" 0.3413

Summary information for training sample of size 1638 (excluding observations with non-positive weight or missing values in d, e, t, r or z variables)
d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators), 
s=split-only numerical, n=split and fit numerical, f=fit-only numerical, 
m=missing-value flag variable, p=periodic variable, w=weight

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>#Codes/Levels/</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>resp6</td>
<td>d</td>
<td>0.000</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>group</td>
<td>r</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>age</td>
<td>s</td>
<td>41.00</td>
<td>75.00</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>educyrs</td>
<td>s</td>
<td>2.000</td>
<td>20.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>fatal</td>
<td>s</td>
<td>11.00</td>
<td>42.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>know</td>
<td>s</td>
<td>1.000</td>
<td>7.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>stage</td>
<td>c</td>
<td></td>
<td></td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

--- Constructed variables ---

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>#Codes/Levels/</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
</tr>
</thead>
<tbody>
<tr>
<td>42</td>
<td>group.DVD</td>
<td>f</td>
<td>0.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>group.Phone</td>
<td>f</td>
<td>0.000</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

Total #cases w/ #missing:
#cases miss. D ord. vals #X-var #N-var #F-var #S-var
1681 43 84 1 0 0 21
#P-var #M-var #B-var #C-var #I-var #R-var
0 0 0 17 0 1

No weight variable in data file
Number of cases used for training: 1638
Number of split variables: 38
Number of dummy variables created: 2
Number of cases excluded due to 0 W or missing D or R variables: 43

Predictive priority (Gi)
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE's for pruned tree: 0.2500

No nodewise interaction tests
Split values for N and S variables based on exhaustive search
Maximum number of split levels: 11
Minimum node sample size: 8
Minimum fraction of cases per treatment at each node: 0.066

Wei-Yin Loh
192
GUIDE manual
11.1 Multiple treatment arms: CAPE data 11 RANDOMIZED TREATMENTS

Top-ranked variables and 1-df chi-squared values at root node
1 0.6775E+01 sf12gh
2 0.5072E+01 know
3 0.3940E+01 incle75k
...
30 0.1110E-03 sf12pf
31 0.1774E-07 sf12mh

Size and CV MSE and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Nodes</th>
<th>Mean MSE</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median MSE</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>125</td>
<td>3.439E-01</td>
<td>9.506E-03</td>
<td>1.064E-02</td>
<td>3.585E-01</td>
<td>1.561E-02</td>
</tr>
<tr>
<td>2</td>
<td>124</td>
<td>3.439E-01</td>
<td>9.506E-03</td>
<td>1.064E-02</td>
<td>3.585E-01</td>
<td>1.561E-02</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>76</td>
<td>13</td>
<td>2.732E-01</td>
<td>6.053E-03</td>
<td>5.569E-03</td>
<td>2.691E-01</td>
<td>6.443E-03</td>
</tr>
<tr>
<td>77</td>
<td>12</td>
<td>2.491E-01</td>
<td>4.721E-03</td>
<td>6.754E-03</td>
<td>2.462E-01</td>
<td>6.768E-03</td>
</tr>
<tr>
<td>78**</td>
<td>5</td>
<td>2.390E-01</td>
<td>3.240E-03</td>
<td>2.264E-03</td>
<td>2.410E-01</td>
<td>3.959E-03</td>
</tr>
<tr>
<td>79++</td>
<td>1</td>
<td>2.414E-01</td>
<td>2.372E-03</td>
<td>5.044E-04</td>
<td>2.410E-01</td>
<td>6.719E-04</td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 5 terminal nodes
0-SE tree based on median is marked with + and has 1 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
** tree same as -- tree
+ tree same as ++ tree
* tree same as ** tree
* tree same as -- tree

Following tree is based on mean CV with naive SE estimate (**)
Number of terminal nodes of final tree: 5
Total number of nodes of final tree: 9
Second best split variable (based on curvature test) at root node is know

Regression tree:

Node 1: sf12gh <= 72.500000
  Node 2: know <= 6.5000000
    Node 4: educyrs <= 15.500000
      Node 8: yearmam <= 3.5000000 or NA
        Node 16: resp6-mean = 0.29976581
        Node 8: yearmam > 3.5000000 or NA
        Node 17: resp6-mean = 0.63793103
      Node 4: educyrs > 15.500000 or NA
        Node 9: resp6-mean = 0.45000000
    Node 2: know > 6.5000000 or NA
      Node 5: resp6-mean = 0.31500000
    Node 1: sf12gh > 72.500000 or NA
      Node 3: resp6-mean = 0.44081633

**********************************************************************
Predictor means below are means of cases with no missing values.
Regression coefficients are computed from the complete cases.

WARNING: p-values below not adjusted for split search. For a bootstrap solution see:

Node 1: Intermediate node
A case goes into Node 2 if sf12gh <= 72.500000
sf12gh mean = 65.921856
Coefficients of least squares regression function:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.3985</td>
<td>18.81</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>group.DVD</td>
<td>-0.7366E-02</td>
<td>-0.2465</td>
<td>0.8054</td>
<td>0.000</td>
<td>0.3309</td>
<td>1.000</td>
</tr>
<tr>
<td>group.Phone</td>
<td>0.2188E-01</td>
<td>0.7378</td>
<td>0.4608</td>
<td>0.000</td>
<td>0.3413</td>
<td>1.000</td>
</tr>
<tr>
<td>resp6 mean</td>
<td>0.403541</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

No truncation of predicted values
### 11.1 Multiple treatment arms: CAPE data

**Node 2: Intermediate node**
A case goes into Node 4 if `know <= 6.500000`

<table>
<thead>
<tr>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.6087154</td>
</tr>
</tbody>
</table>

**Node 4: Intermediate node**
A case goes into Node 8 if `educyrs <= 15.500000`

<table>
<thead>
<tr>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.800853</td>
</tr>
</tbody>
</table>

**Node 8: Intermediate node**
A case goes into Node 16 if `yearmam <= 3.500000`

<table>
<thead>
<tr>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0055249</td>
</tr>
</tbody>
</table>

**Node 16: Terminal node**

**Coefficients of least squares regression functions:**

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.3333</td>
<td>8.279</td>
<td>0.2776E-14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>group.DVD</td>
<td>-0.9843E-01</td>
<td>-1.790</td>
<td>0.7419E-01</td>
<td>0.000</td>
<td>0.3489</td>
<td>1.000</td>
</tr>
<tr>
<td>group.Phone</td>
<td>0.2237E-02</td>
<td>0.4068E-01</td>
<td>0.9676</td>
<td>0.000</td>
<td>0.3489</td>
<td>1.000</td>
</tr>
</tbody>
</table>

res6 mean = 0.299766

No truncation of predicted values

**Node 17: Terminal node**

**Coefficients of least squares regression functions:**

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.5000</td>
<td>6.149</td>
<td>0.1204E-07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>group.DVD</td>
<td>0.1154</td>
<td>1.037</td>
<td>0.3019</td>
<td>0.000</td>
<td>0.3362</td>
<td>1.000</td>
</tr>
<tr>
<td>group.Phone</td>
<td>0.2674</td>
<td>2.458</td>
<td>0.1550E-01</td>
<td>0.000</td>
<td>0.3707</td>
<td>1.000</td>
</tr>
</tbody>
</table>

res6 mean = 0.637931

No truncation of predicted values

**Node 9: Terminal node**

**Coefficients of least squares regression functions:**

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.3788</td>
<td>6.298</td>
<td>0.2840E-08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>group.DVD</td>
<td>0.2366</td>
<td>2.611</td>
<td>0.9889E-02</td>
<td>0.000</td>
<td>0.3250</td>
<td>1.000</td>
</tr>
<tr>
<td>group.Phone</td>
<td>-0.2165E-01</td>
<td>-0.2244</td>
<td>0.8227</td>
<td>0.000</td>
<td>0.2625</td>
<td>1.000</td>
</tr>
</tbody>
</table>

res6 mean = 0.450000

No truncation of predicted values

**Node 5: Terminal node**

**Coefficients of least squares regression functions:**

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.1831</td>
<td>3.417</td>
<td>0.7695E-03</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>group.DVD</td>
<td>0.2883</td>
<td>3.791</td>
<td>0.1993E-03</td>
<td>0.000</td>
<td>0.3500</td>
<td>1.000</td>
</tr>
<tr>
<td>group.Phone</td>
<td>0.1050</td>
<td>1.321</td>
<td>0.1882</td>
<td>0.000</td>
<td>0.2950</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Wei-Yin Loh 195 GUIDE manual
11.2 Censored response: proportional hazards

```
resp6 mean = 0.315000
No truncation of predicted values
----------------------------
Node 3: Terminal node
Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.4895</td>
<td>15.21</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>group.DVD</td>
<td>-0.1101</td>
<td>-2.407</td>
<td>0.1634E-01</td>
<td>0.000</td>
<td>0.3156</td>
<td>1.000</td>
</tr>
<tr>
<td>group.Phone</td>
<td>-0.3832E-01</td>
<td>-0.8659</td>
<td>0.3868</td>
<td>0.000</td>
<td>0.3619</td>
<td>1.000</td>
</tr>
</tbody>
</table>
resp6 mean = 0.440816
No truncation of predicted values
----------------------------
Number of times Li-Martin approximation used = 157
Proportion of variance (R-squared) explained by tree model: 0.0579
```

Observed and fitted values are stored in gi.fit
LaTeX code for tree is in gi.tex
R code is stored in gi.r

The tree has 5 terminal nodes (subgroups) and the results for each terminal node give the treatment effects of DVD and Phone versus Control, which is the first treatment level in alphabetical order. Figure 24 shows the tree diagram.

11.2 Censored response: proportional hazards

We now consider a randomized controlled breast cancer trial where the response variable is a censored survival time (Schmoor et al., 1996). The data are in the file cancerdata.txt; they are included in the TH.data R package (Hothorn, 2017) as well. In the description file cancerdsc.txt below, the treatment variable is hormone therapy, horTh. The variable time is (censored) time to recurrence of cancer and the event indicator event = 1 if the cancer recurred and = 0 if it did not. Ordinal predictor variables may be designated as “n” or “s” (with this option of no linear prognostic control, n variables are automatically changed to s when the program executes). See Loh et al. (2019a, 2016, 2015, 2019c) and Loh and Zhou (2020) for further analysis of the data.

cancerdata.txt
NA
1
1 horTh r
2 age n
3 menostat c
4 tsize n
5 tgrade c

Wei-Yin Loh 196 GUIDE manual
Figure 24: GUIDE v.41.2 0.250-SE least-squares regression tree using Gi option for dependent variable resp6 without adjustment for linear prognostic effects. At each split, an observation goes to the left branch if and only if the condition is satisfied. Sample size (in italics) printed below nodes. resp6 mean for treatment reference level Control followed by treatment effects of levels DVD, Phone (relative to Control) beside nodes. Second best split variable at root node is know.
11.2 Censored response: proportional hazards

11.2.1 Without linear prognostic control

The simplest model only uses the covariates to split the intermediate nodes; terminal nodes are fitted with treatment means.

Input file generation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: ph-gi.in
   Input 1 for model fitting, 2 for importance or DIF scoring,
   3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: ph-gi.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1): 2
Choose type of regression model:
   1=linear, 2=quantile, 3=Poisson, 4=censored response,
   5=multireponse or itemresponse, 6=longitudinal data (with T variables),
   7=binary logistic regression.
   Input choice ([1:7], <cr>=1): 4
   Input 1 for proportional hazards, 2 for restricted mean event time ([1:2], <cr>=1):
Choose complexity of model to use at each node:
   Choose 1 for multiple regression (recommended for prediction)
   Choose 2 for best simple linear in one N or F variable
   Choose 3 for constant fit (recommended for interpretability or if there is an R variable)
   1: multiple linear, 2: best simple linear, 3: constant ([1:3], <cr>=3):
   Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
   Input name of data description file (max 100 characters);
   enclose with matching quotes if it has spaces: cancercdsc.txt
Reading data description file ...
Training sample file: cancercdata.txt
Missing value code: NA
Records in data file start on line 2
R variable present
6 N variables changed to S
Warning: model changed to linear in treatment
11.2  Censored response: proportional hazards

D variable is death
Reading data file ...
Number of records in data file: 686
Length of longest entry in data file: 4
Checking for missing values ...
Finished checking
Assigning integer codes to values of 2 categorical variables
Treatment (R) variable is horTh with values "no" and "yes"
Re-checking data ...
Assigning codes to missing values, if any ...
Data checks complete
Smallest uncensored time: 72.0000
Number of cases dropped due to missing D or T or censored T < smallest uncensored T: 14
Number of complete cases excluding censored T < smallest uncensored T: 672
Number of cases used to compute baseline hazard: 672
Number of cases with D=1 and T >= smallest uncensored: 299
GUIDE will try to create the variables in the description file.
If it is unsuccessful, please create the columns yourself...
Number of dummy variables created: 1
Choose a subgroup identification method:
1 = Prognostic priority (Gs)
2 = Predictive priority (Gi)
Input your choice: ([1:2], <cr>=2):
Creating dummy variables ...
Rereading data ...
Largest uncensored and censored time by horTh
"no" 2456.0000 2563.0000
"yes" 2372.0000 2659.0000
Proportion of training sample for each level of horTh
"no" 0.6399
"yes" 0.3601

Survival time variable in column: 9
Event indicator variable in column: 10
Proportion uncensored among nonmissing T and D variables: .445
Number of cases used for training: 672
Number of split variables: 7
Number of dummy variables created: 1
Finished reading data file
Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):
Input file name to store LaTeX code (use .tex as suffix): ph-gi.tex
You can store the variables and/or values used to split and fit in a file

Wei-Yin Loh

GUIDE manual
Choose 1 to skip this step, 2 to store split and fit variables, 3 to store split variables and their values
Input your choice ([1:3], <cr>=1):
Input 2 to save fitted values and node IDs, 1 otherwise ([1:2], <cr>=2):
Input name of file to store node ID and fitted value of each case: ph-gi.fit
Input 2 to write R function for predicting new cases, 1 otherwise ([1:2], <cr>=2):
Input file name: ph-gi.r
Input rank of top variable to split root node ([1:9], <cr>=1):
Input file is created!
Run GUIDE with the command: guide < ph-gi.in

Results  The contents of ph-gi.out follow.

Regression tree for censored response
Pruning by cross-validation
Data description file: cancerdsc.txt
Training sample file: cancerdata.txt
Missing value code: NA
Records in data file start on line 2
R variable present
6 N variables changed to S
Warning: model changed to linear in treatment
D variable is death
Piecewise linear model
Number of records in data file: 686
Length of longest entry in data file: 4
Treatment (R) variable is horTh with values "no" and "yes"
Number of cases dropped due to missing D or T or censored T < smallest uncensored T: 14
Number of complete cases excluding censored T < smallest uncensored T: 672
Number of cases used to compute baseline hazard: 672
Number of cases with D=1 and T >= smallest uncensored: 299
Number of dummy variables created: 1
Smallest uncensored time: 72.0000
Largest uncensored and censored time by horTh

<table>
<thead>
<tr>
<th>horTh</th>
<th>Uncensored</th>
<th>Censored</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;no&quot;</td>
<td>2456.0000</td>
<td>2563.0000</td>
</tr>
<tr>
<td>&quot;yes&quot;</td>
<td>2372.0000</td>
<td>2659.0000</td>
</tr>
</tbody>
</table>

Proportion of training sample for each level of horTh
"no" 0.6399
"yes" 0.3601

Summary information for training sample of size 672 (excluding observations with non-positive weight or missing values in d, e, t, r or z variables)
d=dependent, b=split and fit cat variable using indicator variables, c=split-only categorical, i=fit-only categorical (via indicators),
11.2 Censored response: proportional hazards

s=split-only numerical, n=split and fit numerical, f=fit-only numerical, 
m=missing-value flag variable, p=periodic variable, w=weight, 
t=survival time variable

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>horTh</td>
<td>r</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>age</td>
<td>s</td>
<td>21.00</td>
<td>80.00</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>menostat</td>
<td>c</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>tsize</td>
<td>s</td>
<td>3.000</td>
<td>120.0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>tgrade</td>
<td>s</td>
<td>1.000</td>
<td>3.000</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>pnodes</td>
<td>s</td>
<td>1.000</td>
<td>51.00</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>progres</td>
<td>s</td>
<td>0.000</td>
<td>2380.</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>estrec</td>
<td>s</td>
<td>0.000</td>
<td>1144.</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>time</td>
<td>t</td>
<td>72.00</td>
<td>2659.</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>death</td>
<td>d</td>
<td>0.000</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

Constructive variables:

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>lnbasehaz</td>
<td>z</td>
<td>-6.510</td>
<td>0.5887E-01</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>horTh.yes</td>
<td>f</td>
<td>0.000</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

Total #cases w/ #missing:

<table>
<thead>
<tr>
<th>#cases</th>
<th>miss. D</th>
<th>ord. vals</th>
<th>#X-var</th>
<th>#N-var</th>
<th>#F-var</th>
<th>#S-var</th>
<th>#P-var</th>
<th>#M-var</th>
<th>#B-var</th>
<th>#C-var</th>
<th>#I-var</th>
<th>#R-var</th>
</tr>
</thead>
<tbody>
<tr>
<td>686</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Survival time variable in column: 9
Event indicator variable in column: 10
Proportion uncensored among nonmissing T and D variables: 0.445
Number of cases used for training: 672
Number of split variables: 7
Number of dummy variables created: 1

Constant fitted to cases with missing values in regressor variables
Predictive priority (Gi)
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE’s for pruned tree: 0.2500

No nodewise interaction tests
Fraction of cases used for splitting each node: 1.0000
Maximum number of split levels: 10
Minimum node sample size: 6
Minimum fraction of cases per treatment at each node: 0.072
Number of iterations for fitting: 20
Top-ranked variables and 1-df chi-squared values at root node

1 0.2101E+01 progres
11.2 Censored response: proportional hazards

```
2 0.1669E+01  estrec
3 0.1108E+01  tsize
4 0.3557E+00  pnodes
5 0.2413E+00  tgrade
6 0.2057E-01  menostat
7 0.1879E-02  age
```

Size and CV Loss and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Tnodes</th>
<th>Mean Loss</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median Loss</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>48</td>
<td>1.739E+00</td>
<td>8.406E-02</td>
<td>6.834E-02</td>
<td>1.706E+00</td>
<td>7.329E-02</td>
</tr>
<tr>
<td>2</td>
<td>47</td>
<td>1.737E+00</td>
<td>8.408E-02</td>
<td>6.866E-02</td>
<td>1.697E+00</td>
<td>7.379E-02</td>
</tr>
<tr>
<td>:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29**</td>
<td>4</td>
<td>1.461E+00</td>
<td>6.040E-02</td>
<td>4.355E-02</td>
<td>1.443E+00</td>
<td>4.585E-02</td>
</tr>
<tr>
<td>30**</td>
<td>2</td>
<td>1.398E+00</td>
<td>5.064E-02</td>
<td>1.949E-02</td>
<td>1.400E+00</td>
<td>2.803E-02</td>
</tr>
<tr>
<td>31</td>
<td>1</td>
<td>1.435E+00</td>
<td>5.100E-02</td>
<td>1.066E-02</td>
<td>1.446E+00</td>
<td>1.482E-02</td>
</tr>
</tbody>
</table>

- SE tree based on mean is marked with * and has 2 terminal nodes
- Selected-SE tree based on mean using naive SE is marked with **
- Selected-SE tree based on mean using bootstrap SE is marked with --
- Selected-SE tree based on median and bootstrap SE is marked with ++

* tree, ** tree, + tree, and ++ tree all the same

Following tree is based on mean CV with naive SE estimate (**)

Structure of final tree. Each terminal node is marked with a T.

Cases fit give the number of cases used to fit node

<table>
<thead>
<tr>
<th>Node</th>
<th>Total Cases</th>
<th>Matrix Median Node Split</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>label cases</td>
<td>rank survtime deviance variable</td>
</tr>
<tr>
<td>1</td>
<td>672 672</td>
<td>1 1.807E+03 estrec</td>
</tr>
<tr>
<td>2T</td>
<td>274 274</td>
<td>1 1.140E+03 estrec</td>
</tr>
<tr>
<td>3T</td>
<td>398 398</td>
<td>1 2.286E+03 menostat</td>
</tr>
</tbody>
</table>

Number of terminal nodes of final tree: 2
Total number of nodes of final tree: 3
Second best split variable (based on curvature test) at root node is estrec

Regression tree:

Node 1: progres <= 21.500000
    Node 2: Median survival time = 1140.0000
Node 1: progres > 21.500000 or NA
    Node 3: Median survival time = 2286.0000
WARNING: p-values below not adjusted for split search. For a bootstrap solution see:


Node 1: Intermediate node
A case goes into Node 2 if progrec <= 21.500000
progrec mean = 110.91518
Coefficients of log-relative hazard function (relative to baseline hazard):

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>horTh.yes</td>
<td>-0.3654</td>
<td>-2.933</td>
<td>0.3471E-02</td>
<td>0.000</td>
<td>0.3601</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Predicted log-relative hazard = -0.30206062E-2

Node 2: Terminal node
Coefficients of log-relative hazard function (relative to baseline hazard):

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.3729</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>horTh.yes</td>
<td>-0.1140</td>
<td>-0.6871</td>
<td>0.4926</td>
<td>0.000</td>
<td>0.3613</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Predicted log-relative hazard = 0.45682185

Node 3: Terminal node
Coefficients of log-relative hazard function (relative to baseline hazard):

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.2596</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>horTh.yes</td>
<td>-0.6453</td>
<td>-3.375</td>
<td>0.8098E-03</td>
<td>0.000</td>
<td>0.3593</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Predicted log-relative hazard = -0.34487497

Observed and fitted values are stored in ph-gi.fit
LaTeX code for tree is in ph-gi.tex
R code is stored in ph-gi.r

Let $\lambda(u, x)$ denote the hazard function at time $u$ and predictor values $x$ and let $\lambda_0(u)$ denote the baseline hazard function. The results in ph-gi.out show that the...
11.2 Censored response: proportional hazards

Figure 25: GUIDE v.41.1 0.250-SE proportional hazards regression tree using Gi option for time and event indicator death without adjustment for linear prognostic effects. (constant fitted to incomplete cases in terminal nodes). At each split, an observation goes to the left branch if and only if the condition is satisfied. Treatment horTh hazard ratio of level yes to level no beside nodes. Sample size (in italics), median survival time, and proportion of horTh = yes printed below nodes. Terminal nodes with treatment hazard ratio above and below 0.694 (ratio at root node) are colored orange and skyblue respectively. Second best split variable at root node is estrec.

Node 2

Node 3

Figure 26: Estimated survival probability functions for breast cancer data
fitted proportional hazards model is

\[ \lambda(u, x) = \lambda_0(u)[\exp(\hat{\beta}_1 + \hat{\gamma}_1 I(\text{horTh} = \text{yes}))I(\text{progres} \leq 21.5) + \exp(\hat{\beta}_2 + \hat{\gamma}_2 I(\text{horTh} = \text{yes}))I(\text{progres} > 21.5)] \]

with \( \hat{\beta}_1 = 0.37292, \hat{\gamma}_1 = -0.11404, \hat{\beta}_2 = -0.25964, \) and \( \hat{\gamma}_2 = -0.64531. \)

Figure 25 shows the tree diagram. The numbers beside each terminal node are relative hazards of \( \text{horTh} = \text{yes} \) versus \( \text{no} \), namely, \( \exp(\hat{\gamma}_1) = \exp(-0.11404) = 0.892223 \) for node 2 and \( \exp(\hat{\gamma}_2) = \exp(-0.64531) = 0.5244999 \) for node 3. Figure 26 shows Kaplan-Meier survival functions of the data in the terminal nodes. The plots are produced by the following R code.

```r
library(survival)
z <- read.table("cancerdata.txt",header=TRUE)
leg.txt <- c("horTh = yes","horTh = no")
leg.col <- c("red","blue")
leg.lty <- 1:2
xr <- range(z$time)
zg <- read.table("ph-gi.fit",header=TRUE)
nodes <- zg$node
uniq.gp <- unique(sort(nodes))
plotted <- FALSE
for(g in uniq.gp){
gp <- nodes == g
y <- z$time[g]
stat <- z$death[g]
treat <- z$horTh[g]
fit <- survfit(Surv(y,stat) ~ treat, conf.type="none")
if(plotted){
    plot(fit,xlim=xr,mark.time=FALSE,xlab="",ylab="",col=c("blue","red"),lwd=2)
} else {
    plot(fit,xlim=xr,mark.time=FALSE,xlab="",ylab="Survival probability",
         col=c("blue","red"),lwd=2)
    plotted <- TRUE
}
title(paste("Node",g))
legend("bottomleft",legend=leg.txt,lty=1,col=leg.col,lwd=2)
}
```

**Estimated relative risks and survival probabilities** The file `ph-gi.fit` gives the terminal node number, observed survival time, event indicator (\( y=\text{uncensored}, n=\text{censored} \)), log baseline cumulative hazard, survival probability, median survival time, and treatment effect (regression coefficient of treatment indicator) of each
observation in the training sample (cancerdata.txt). The results for the first few observations are shown below.

<table>
<thead>
<tr>
<th>train</th>
<th>node</th>
<th>observed</th>
<th>event</th>
<th>logbasecumhaz</th>
<th>survivalprob</th>
<th>mediansurvtime</th>
<th>horTh.yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>3</td>
<td>y</td>
<td>1814.00</td>
<td>-0.335623</td>
<td>0.576131</td>
<td>2286.00</td>
<td>-0.645311</td>
</tr>
<tr>
<td>y</td>
<td>3</td>
<td>y</td>
<td>2018.00</td>
<td>-0.210308</td>
<td>0.720485</td>
<td>2286.00</td>
<td>-0.645311</td>
</tr>
<tr>
<td>y</td>
<td>3</td>
<td>y</td>
<td>712.00</td>
<td>-1.28452</td>
<td>0.894065</td>
<td>2286.00</td>
<td>-0.645311</td>
</tr>
<tr>
<td>y</td>
<td>3</td>
<td>y</td>
<td>1807.00</td>
<td>-0.358191</td>
<td>0.753697</td>
<td>2286.00</td>
<td>-0.645311</td>
</tr>
<tr>
<td>y</td>
<td>3</td>
<td>y</td>
<td>772.00</td>
<td>-1.16232</td>
<td>0.785652</td>
<td>2286.00</td>
<td>-0.645311</td>
</tr>
<tr>
<td>y</td>
<td>2</td>
<td>y</td>
<td>448.00</td>
<td>-2.08322</td>
<td>0.834592</td>
<td>1140.00</td>
<td>-0.114042</td>
</tr>
<tr>
<td>y</td>
<td>3</td>
<td>n</td>
<td>2172.00</td>
<td>-0.121866</td>
<td>0.698971</td>
<td>2286.00</td>
<td>-0.645311</td>
</tr>
</tbody>
</table>

11.2.2 Simple linear prognostic control

To reduce or eliminate confounding between treatment and covariate variables, it may be desirable to adjust for the effects of the latter by fitting a regression model that allows for the linear effects of one or more prognostic variables in each node (Loh et al., 2019c). This is done by choosing the “simple linear” or the “multiple linear” option and specifying each potential linear predictor as “n” in the description file (no change is needed in cancerdsc.txt). First we show how to choose the simple linear model, where a single prognostic variable is used as regressor in each node. There are two options: the Gi (default) option is more sensitive to detecting predictive variables while the Gs option is equally sensitive to detecting prognostic variables—see Loh et al. (2015) for definitions.

Input file generation for Gi method

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: lin-gi.in
   Input 1 for model fitting, 2 for importance or DIF scoring, 3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: lin-gi.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1): 2
   Choose type of regression model:
   1=linear, 2=quantile, 3=Poisson, 4=censored response, 5=multiresponse or itemresponse, 6=longitudinal data (with T variables), 7=multivariate logistic regression.
   Input choice ([1:7], <cr>=1): 4
   Input 1 for proportional hazards, 2 for restricted mean event time ([1:2], <cr>=1):
   Choose complexity of model to use at each node:
11.2  Censored response: proportional hazards

RANDOMIZED TREATMENTS

Choose 1 for multiple regression (recommended for prediction)
Choose 2 for best simple linear in one N or F variable
Choose 3 for constant fit (recommended for interpretability or if there is an R variable)
1: multiple linear, 2: best simple linear, 3: constant ([1:3], <cr>=3): 2
Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
Input name of data description file (max 100 characters);
enclose with matching quotes if it has spaces: cancerdsc.txt

Reading data description file ...
Training sample file: cancerdata.txt
Missing value code: NA
Records in data file start on line 2
R variable present
D variable is death
Reading data file ...
Number of records in data file: 686
Length of longest entry in data file: 4
Checking for missing values ...
Finished checking
Assigning integer codes to values of 2 categorical variables
Treatment (R) variable is horTh with values "no" and "yes"
Re-checking data ...
Assigning codes to missing values, if any ...
Data checks complete
Smallest uncensored time: 72.0000
Number of cases dropped due to missing D or T or censored T < smallest uncensored T: 14
Number of complete cases excluding censored T < smallest uncensored T: 672
Number of cases used to compute baseline hazard: 672
Number of cases with D=1 and T >= smallest uncensored: 299
GUIDE will try to create the variables in the description file.
If it is unsuccessful, please create the columns yourself...
Number of dummy variables created: 1
Choose a subgroup identification method:
1 = Prognostic priority (Gs)
2 = Predictive priority (Gi)
Input your choice: ([1:2], <cr>=2):
Creating dummy variables ...
Rereading data ...
Largest uncensored and censored time by horTh
"no" 2456.0000 2563.0000
"yes" 2372.0000 2659.0000
Proportion of training sample for each level of horTh
"no" 0.6399
"yes" 0.3601

Total  #cases w/  #missing
#cases miss. D ord. vals #X-var #N-var #F-var #S-var
686 0 0 0 6 0 0 0

Wei-Yin Loh 207  GUIDE manual
Results for Gi method  The following output shows that the pruned tree is trivial with no splits and that the variable pnodes is the best simple linear predictor.

Regression tree for censored response
No truncation of predicted values
Pruning by cross-validation
Data description file: cancerdsc.txt
Training sample file: cancerdata.txt
Missing value code: NA
Records in data file start on line 2
R variable present
D variable is death
Piecewise simple linear or constant model
Powers are dropped if they are not significant at level 1.0000
Number of records in data file: 686
Length of longest entry in data file: 4
Treatment (R) variable is horTh with values "no" and "yes"
Number of cases dropped due to missing D or T or censored T < smallest uncensored T: 14

Wei-Yin Loh  208  GUIDE manual
Number of complete cases excluding censored T < smallest uncensored T: 672
Number of cases used to compute baseline hazard: 672
Number of cases with D=1 and T >= smallest uncensored: 299
Number of dummy variables created: 1
Smallest uncensored time: 72.0000
Largest uncensored and censored time by horTh

<table>
<thead>
<tr>
<th>horTh</th>
<th>Uncensored</th>
<th>Censored</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;no&quot;</td>
<td>2456.0000</td>
<td>2563.0000</td>
</tr>
<tr>
<td>&quot;yes&quot;</td>
<td>2372.0000</td>
<td>2659.0000</td>
</tr>
</tbody>
</table>

Proportion of training sample for each level of horTh

<table>
<thead>
<tr>
<th>horTh</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;no&quot;</td>
<td>0.6399</td>
</tr>
<tr>
<td>&quot;yes&quot;</td>
<td>0.3601</td>
</tr>
</tbody>
</table>

Survival time variable in column: 9
Event indicator variable in column: 10
Proportion uncensored among nonmissing T and D variables: 0.445
Number of cases used for training: 672
Number of split variables: 7
Number of dummy variables created: 1

Warning: missing regressor values imputed with node means
Predictive priority (Gi)
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE’s for pruned tree: 0.2500

No nodewise interaction tests
Fraction of cases used for splitting each node: 1.0000
Maximum number of split levels: 10
Minimum node sample size: 7
Minimum fraction of cases per treatment at each node: 0.072
Number of iterations for fitting: 20

Top-ranked variables and 1-df chi-squared values at root node
   1 0.3130E+01 estrec
   2 0.1672E+01 progrec
   3 0.1137E+01 tsize
   4 0.3983E+00 pnodes
   5 0.1718E+00 tgrade
   6 0.9820E-01 menostat
   7 0.2054E-04 age

Size and CV Loss and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Tnodes</th>
<th>Mean Loss</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median Loss</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>43</td>
<td>1.247E+07</td>
<td>1.219E+07</td>
<td>1.214E+07</td>
<td>7.263E+00</td>
<td>3.919E+06</td>
</tr>
<tr>
<td>2</td>
<td>42</td>
<td>1.247E+07</td>
<td>1.219E+07</td>
<td>1.214E+07</td>
<td>7.266E+00</td>
<td>3.919E+06</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>8</td>
<td>2.741E+05</td>
<td>2.739E+05</td>
<td>2.591E+05</td>
<td>2.918E+00</td>
<td>5.727E-01</td>
</tr>
<tr>
<td>20</td>
<td>6</td>
<td>2.741E+05</td>
<td>2.739E+05</td>
<td>2.591E+05</td>
<td>1.542E+00</td>
<td>2.450E-01</td>
</tr>
<tr>
<td>21++</td>
<td>2</td>
<td>1.370E+00</td>
<td>7.295E-02</td>
<td>5.276E-02</td>
<td>1.320E+00</td>
<td>3.197E-02</td>
</tr>
<tr>
<td>22**</td>
<td>1</td>
<td>1.355E+00</td>
<td>5.363E-02</td>
<td>2.719E-02</td>
<td>1.330E+00</td>
<td>2.698E-02</td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 1 terminal node
0-SE tree based on median is marked with + and has 2 terminal node
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
++ tree same as -- tree
+ tree same as ++ tree
* tree same as ** tree
* tree same as -- tree

Following tree is based on mean CV with naive SE estimate (**)
11.2  Censored response: proportional hazards

Randomized Treatments

Structure of final tree. Each terminal node is marked with a T.

Cases fit give the number of cases used to fit node
Deviance is mean residual deviance for all cases in node

<table>
<thead>
<tr>
<th>Node</th>
<th>Total Cases</th>
<th>Matrix Rank</th>
<th>Median</th>
<th>Node Split</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1T</td>
<td>672</td>
<td>672</td>
<td>3</td>
<td>1.807E+03</td>
<td>1.343E+00</td>
</tr>
</tbody>
</table>

Best split at root node is estrec <= 4.5000

Number of terminal nodes of final tree: 1
Total number of nodes of final tree: 1
Best split variable (based on curvature test) at root node is estrec

Regression tree:

Node 1: Median survival time = 1807.0000

***************************************************************
Node 1: Terminal node

Coefficients of log-relative hazard function (relative to baseline hazard):

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pnodes</td>
<td>0.5630E-01</td>
<td>8.575</td>
<td>0.000</td>
<td>1.000</td>
<td>4.987</td>
<td>51.00</td>
</tr>
<tr>
<td>horTh.yes</td>
<td>-0.3465</td>
<td>-2.778</td>
<td>0.5627E-02</td>
<td>0.000</td>
<td>0.3601</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Observed and fitted values are stored in lin-gi.fit
Regressor names and coefficients are stored in lin-gi.reg
LaTeX code for tree is in lin-gi.tex
R code is stored in lin-gi.r

The file lin-gi.reg reports the selected regressor in each terminal node of the tree (there is only one node here):

node bestvar
1 pnodes

Input file generation for Gs method

0. Read the warranty disclaimer
1. Create a GUIDE input file
Input your choice: 1
Name of batch input file: lin-gs.in
Input 1 for model fitting, 2 for importance or DIF scoring,
11.2 Censored response: proportional hazards

for data conversion ([1:3], <cr>=1):
Name of batch output file: lin-gs.out
Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
Input 1 for classification, 2 for regression, 3 for propensity score grouping
Input your choice ([1:3], <cr>=1): 2
Choose type of regression model:
1=linear, 2=quantile, 3=Poisson, 4=censored response,
5=multiresponse or itemresponse, 6=longitudinal data (with T variables),
7=binary logistic regression.
Input choice ([1:7], <cr>=1): 4
Input 1 for proportional hazards, 2 for restricted mean event time ([1:2], <cr>=1):
Choose complexity of model to use at each node:
Choose 1 for multiple regression (recommended for prediction)
Choose 2 for best simple linear in one N or F variable
Choose 3 for constant fit (recommended for interpretability or if there is an R variable)
1: multiple linear, 2: best simple linear, 3: constant ([1:3], <cr>=3): 2
Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
Input name of data description file (max 100 characters);
enclose with matching quotes if it has spaces: cancerdsc.txt
Reading data description file ...
Training sample file: cancerdata.txt
Missing value code: NA
Records in data file start on line 2
R variable present
D variable is death
Reading data file ...
Number of records in data file: 686
Length of longest entry in data file: 4
Checking for missing values ...
Finished checking
Assigning integer codes to values of 2 categorical variables
Treatment (R) variable is horTh with values "no" and "yes"
Re-checking data ...
Assigning codes to missing values, if any ...
Data checks complete
Smallest uncensored time: 72.0000
Number of cases dropped due to missing D or T or censored T < smallest uncensored T: 14
Number of complete cases excluding censored T < smallest uncensored T: 672
Number of cases used to compute baseline hazard: 672
Number of cases with D=1 and T >= smallest uncensored: 299
GUIDE will try to create the variables in the description file.
If it is unsuccessful, please create the columns yourself...
Number of dummy variables created: 1
Choose a subgroup identification method:
1 = Prognostic priority (Gs)
2 = Predictive priority (Gi)
Results for Gs method  The Gs method gives a tree with three terminal nodes.

Regression tree for censored response
No truncation of predicted values
Pruning by cross-validation
Data description file: cancerdsc.txt
Training sample file: cancerdata.txt

Wei-Yin Loh  213  GUIDE manual
11.2 Censored response: proportional hazards

Missing value code: NA
Records in data file start on line 2
R variable present
D variable is death
Piecewise simple linear or constant model
Powers are dropped if they are not significant at level 1.0000
Number of records in data file: 686
Length of longest entry in data file: 4
Treatment (R) variable is horTh with values "no" and "yes"
Number of cases dropped due to missing D or T or censored T < smallest uncensored T: 14
Number of complete cases excluding censored T < smallest uncensored T: 672
Number of cases used to compute baseline hazard: 672
Number of cases with D=1 and T \( \geq \) smallest uncensored: 299
Number of dummy variables created: 1
Smallest uncensored time: 72.0000
Largest uncensored and censored time by horTh

<table>
<thead>
<tr>
<th>horTh</th>
<th>Uncensored</th>
<th>Censored</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;no&quot;</td>
<td>2456.0000</td>
<td>2563.0000</td>
</tr>
<tr>
<td>&quot;yes&quot;</td>
<td>2372.0000</td>
<td>2659.0000</td>
</tr>
</tbody>
</table>

Proportion of training sample for each level of horTh
"no" 0.6399
"yes" 0.3601

Summary information for training sample of size 672 (excluding observations with non-positive weight or missing values in d, e, t, r or z variables)
d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators),
s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
m=missing-value flag variable, p=periodic variable, w=weight,
t=survival time variable

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>horTh</td>
<td>r</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>age</td>
<td>n</td>
<td>21.00</td>
<td>80.00</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>menostat</td>
<td>c</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>tsize</td>
<td>n</td>
<td>3.000</td>
<td>120.0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>tgrade</td>
<td>n</td>
<td>1.000</td>
<td>3.000</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>pnodes</td>
<td>n</td>
<td>1.000</td>
<td>51.00</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>progrec</td>
<td>n</td>
<td>0.000</td>
<td>2380.</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>estrec</td>
<td>n</td>
<td>0.000</td>
<td>1144.</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>time</td>
<td>t</td>
<td>72.00</td>
<td>2659.</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>death</td>
<td>d</td>
<td>0.000</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

============================ Constructed variables =============================

11 lnbasehaz z -6.510 0.5887E-01
12 horTh.yes f 0.000 1.000

Wei-Yin Loh 214 GUIDE manual
11.2 Censored response: proportional hazards

Randomized Treatments

<table>
<thead>
<tr>
<th>Total cases w/</th>
<th>cases miss. D ord. vals</th>
<th>#X-var</th>
<th>#N-var</th>
<th>#F-var</th>
<th>#S-var</th>
<th>#P-var</th>
<th>#M-var</th>
<th>#B-var</th>
<th>#C-var</th>
<th>#I-var</th>
<th>#R-var</th>
</tr>
</thead>
<tbody>
<tr>
<td>686</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Survival time variable in column: 9
Event indicator variable in column: 10
Proportion uncensored among nonmissing T and D variables: 0.445
Number of cases used for training: 672
Number of split variables: 7
Number of dummy variables created: 1

Warning: missing regressor values imputed with node means
Prognostic priority (Gs)
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE's for pruned tree: 0.2500

No nodewise interaction tests
Fraction of cases used for splitting each node: 1.0000
Maximum number of split levels: 10
Minimum node sample size: 7
Minimum fraction of cases per treatment at each node: 0.072
Number of iterations for fitting: 20
Top-ranked variables and 1-df chi-squared values at root node
1 0.2695E+02 pnodes
2 0.1812E+02 progrec
3 0.8046E+01 estrec
4 0.3781E+01 tgrade
5 0.8274E+00 menostat
6 0.5154E+00 tsize
7 0.3349E+00 age

Size and CV Loss and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Nodes</th>
<th>Mean Loss</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median Loss</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>45</td>
<td>9.913E+03</td>
<td>9.901E+03</td>
<td>9.080E+03</td>
<td>3.361E+00</td>
<td>7.785E-01</td>
</tr>
<tr>
<td>2</td>
<td>44</td>
<td>9.913E+03</td>
<td>9.901E+03</td>
<td>9.080E+03</td>
<td>3.092E+00</td>
<td>8.253E-01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>7</td>
<td>1.956E+00</td>
<td>3.378E-01</td>
<td>2.894E-01</td>
<td>1.630E+00</td>
<td>1.805E-01</td>
</tr>
<tr>
<td>20</td>
<td>4</td>
<td>1.432E+00</td>
<td>6.770E-02</td>
<td>5.670E-02</td>
<td>1.424E+00</td>
<td>7.438E-02</td>
</tr>
<tr>
<td>21**</td>
<td>3</td>
<td>1.336E+00</td>
<td>5.196E-02</td>
<td>3.403E-02</td>
<td>1.289E+00</td>
<td>3.960E-02</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>1.362E+00</td>
<td>5.631E-02</td>
<td>3.638E-02</td>
<td>1.314E+00</td>
<td>5.650E-02</td>
</tr>
<tr>
<td>23</td>
<td>1</td>
<td>1.383E+00</td>
<td>5.502E-02</td>
<td>2.787E-02</td>
<td>1.359E+00</td>
<td>2.776E-02</td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 3 terminal nodes
11.2  Censored response: proportional hazard

Randomized treatments

0-SE tree based on median is marked with + and has 3 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
* tree, ** tree, + tree, and ++ tree all the same

Following tree is based on mean CV with naive SE estimate (**)

Structure of final tree. Each terminal node is marked with a T.

Cases fit give the number of cases used to fit node
Deviance is mean residual deviance for all cases in node

<table>
<thead>
<tr>
<th>Node</th>
<th>Total Cases</th>
<th>Matrix</th>
<th>Median Node Split</th>
<th>Cases</th>
<th>Rank</th>
<th>Survtime</th>
<th>Deviance</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>672</td>
<td>672</td>
<td>3 1.807E+03 1.371E+00</td>
<td>pnodes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>370</td>
<td>370</td>
<td>3 2.659E+03+ 1.092E+00</td>
<td>age</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4T</td>
<td>142</td>
<td>142</td>
<td>3 2.563E+03+ 9.548E-01</td>
<td>tsize</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5T</td>
<td>228</td>
<td>228</td>
<td>3 2.030E+03 1.044E+00</td>
<td>tgrade</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3T</td>
<td>302</td>
<td>302</td>
<td>3 9.830E+02 1.552E+00</td>
<td>progres</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Number of terminal nodes of final tree: 3
Total number of nodes of final tree: 5
Second best split variable (based on curvature test) at root node is progres

Regression tree:

Node 1: pnodes <= 3.5000000
   Node 2: age <= 49.500000
      Node 4: Median survival time = 2563.0000+
   Node 2: age > 49.500000 or NA
      Node 5: Median survival time = 2030.0000
Node 1: pnodes > 3.5000000 or NA
   Node 3: Median survival time = 983.00000

*****************************************************************************

Warning: p-values below not adjusted for split search. For a bootstrap solution see:


Wei-Yin Loh 216 GUIDE manual
Node 1: Intermediate node
A case goes into Node 2 if \( p\text{nodes} \leq 3.500000 \)
\( p\text{nodes mean} = 4.9866071 \)

Coefficients of log-relative hazard function (relative to baseline hazard):

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p\text{nodes} )</td>
<td>0.5725E-01</td>
<td>8.744</td>
<td>0.000</td>
<td>1.000</td>
<td>4.987</td>
<td>51.00</td>
</tr>
<tr>
<td>( \text{horTh.yes} )</td>
<td>-0.3528</td>
<td>-2.828</td>
<td>0.4823E-02</td>
<td>0.000</td>
<td>0.3601</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Node 2: Intermediate node
A case goes into Node 4 if \( \text{age} \leq 49.500000 \)
\( \text{age mean} = 53.235135 \)

Node 4: Terminal node
Coefficients of log-relative hazard function (relative to baseline hazard):

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>5.162</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \text{age} )</td>
<td>-0.1344</td>
<td>-5.463</td>
<td>0.2096E-06</td>
<td>21.00</td>
<td>43.00</td>
<td>49.00</td>
</tr>
<tr>
<td>( \text{horTh.yes} )</td>
<td>-0.7981</td>
<td>-1.502</td>
<td>0.1353</td>
<td>0.000</td>
<td>0.1690</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Node 5: Terminal node
Coefficients of log-relative hazard function (relative to baseline hazard):

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.3737</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \text{progrec} )</td>
<td>-0.3152E-02</td>
<td>-2.547</td>
<td>0.1152E-01</td>
<td>0.000</td>
<td>112.1</td>
<td>1490.</td>
</tr>
<tr>
<td>( \text{horTh.yes} )</td>
<td>-0.6723</td>
<td>-2.877</td>
<td>0.4400E-02</td>
<td>0.000</td>
<td>0.4474</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Node 3: Terminal node
Coefficients of log-relative hazard function (relative to baseline hazard):

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1.039</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \text{progrec} )</td>
<td>-0.2870E-02</td>
<td>-4.036</td>
<td>0.6925E-04</td>
<td>0.000</td>
<td>105.2</td>
<td>2380.</td>
</tr>
<tr>
<td>( \text{horTh.yes} )</td>
<td>-0.3303</td>
<td>-2.112</td>
<td>0.3549E-01</td>
<td>0.000</td>
<td>0.3841</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Observed and fitted values are stored in lin-gs.fit
Regressor names and coefficients are stored in lin-gs.reg
LaTeX code for tree is in lin-gs.tex
R code is stored in lin-gs.r

The tree is shown in Figure 27. It does not display the linear predictor selected at each terminal node. This information is given in the file lin-gs.out or, more conveniently, in tabular form in lin-gs.reg as shown below.
Figure 27: GUIDE v.41.2 0.250-SE proportional hazards regression tree using Gs option for time and event indicator death with adjustment for simple linear prognostic effects (missing regressor values imputed with node means). At each split, an observation goes to the left branch if and only if the condition is satisfied. Sample size (in italics), median survival time, and proportion of horTh = yes printed below nodes. Treatment horTh hazard ratio of level yes to no beside nodes. Terminal nodes with treatment hazard ratio above and below 0.703 (ratio at root node) are colored yellow and skyblue respectively. Second best split variable at root node is progres.
11.3 Censored response: restricted mean

Besides a proportional hazards tree, GUIDE can also fit a tree to estimate the restricted mean survival time in each node (Chen and Tsiatis, 2001; Tian et al., 2014). This section shows how this is carried out. The time restriction may be changed by the user during when the input file is created.

11.3.1 Without linear prognostic control

The piecewise-constant Gi tree has no splits when the restricted mean option is chosen.

Input file generation for Gi method

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: rest-gi.in
   Input 1 for model fitting, 2 for importance or DIF scoring,
   3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: rest-gi.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1): 2
   Choose type of regression model:
   1=linear, 2=quantile, 3=Poisson, 4=censored response,
   5=multiresponse or itemresponse, 6=longitudinal data (with T variables),
   7=binary logistic regression.
   Input choice ([1:7], <cr>=1): 4
   Input 1 for proportional hazards, 2 for restricted mean event time ([1:2], <cr>=1): 2
   Choose complexity of model to use at each node:
   Choose 0 for stepwise linear regression (recommended for prediction)
   Choose 1 for multiple regression
   Choose 2 for best simple polynomial in one N or F variable
   Choose 3 for constant fit (recommended for interpretability or if there is an R variable)
   0: stepwise linear, 1: multiple linear, 2: best simple polynomial, 3: constant,
   4: best simple stepwise ANCOVA ([0:4], <cr>=3):
   Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
   Input name of data description file (max 100 characters);
enclose with matching quotes if it has spaces: cancerdsc.txt

Reading data description file ...
Training sample file: cancerdata.txt
Missing value code: NA
Records in data file start on line 2
R variable present
6 N variables changed to S
Warning: model changed to linear in treatment
D variable is death
Reading data file ...
Number of records in data file: 686
Length of longest entry in data file: 4
Checking for missing values ...
Finished checking
Assigning integer codes to values of 2 categorical variables
Treatment (R) variable is horTh with values "no" and "yes"
Re-checking data ...
Assigning codes to missing values, if any ...
Data checks complete
GUIDE will try to create the variables in the description file.
If it is unsuccessful, please create the columns yourself...
Number of dummy variables created: 1
Choose a subgroup identification method:
1 = Prognostic priority (Gs)
2 = Predictive priority (Gi)
Input your choice: ([1:2], <cr>=2):
Creating dummy variables ...
Rereading data ...
Largest uncensored and censored time by horTh
"no" 2456.0000 2563.0000
"yes" 2372.0000 2659.0000
Smallest observed uncensored time is 72.0000
Largest observed censored or uncensored time is 2659.0000
Input restriction on event time ([72.00:2659.00], <cr>=1222.00):
Proportion of training sample for each level of horTh
"no" 0.6360
"yes" 0.3640
No weight variable in data file
Number of cases used for training: 533
Number of split variables: 7
Number of dummy variables created: 1
Results for Gi method

Restricted mean event time regression tree
Pruning by cross-validation
Data description file: cancerdsc.txt
Training sample file: cancerdata.txt
Missing value code: NA
Records in data file start on line 2
R variable present
6 N variables changed to S
Warning: model changed to linear in treatment
D variable is death
Piecewise linear model
Number of records in data file: 686
Length of longest entry in data file: 4
Treatment (R) variable is horTh with values "no" and "yes"
Number of dummy variables created: 1
Smallest uncensored time: 72.0000
Largest uncensored and censored time by horTh
\[
\begin{array}{lll}
\text{horTh} & \text{Uncensored} & \text{Censored} \\
\text{"no"} & 2456.0000 & 2563.0000 \\
\text{"yes"} & 2372.0000 & 2659.0000 \\
\end{array}
\]
Interval for restricted mean event time is from 0 to 1222.
Proportion of training sample for each level of horTh
\[
\begin{array}{l}
\text{"no"} & 0.6360 \\
\text{"yes"} & 0.3640 \\
\end{array}
\]

Summary information for training sample of size 533 (excluding observations with non-positive weight or missing values in d, e, t, r or z variables)
d=dependent, b=split and fit cat variable using indicator variables,
11.3 Censored response: restricted mean

RANDOMIZED TREATMENTS

c=split-only categorical, i=fit-only categorical (via indicators),
s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
m=missing-value flag variable, p=periodic variable, w=weight

<table>
<thead>
<tr>
<th># Codes/</th>
<th>Levels/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Column</td>
<td>Name</td>
</tr>
<tr>
<td>1 horTh</td>
<td>r</td>
</tr>
<tr>
<td>2 age</td>
<td>s</td>
</tr>
<tr>
<td>3 menostat</td>
<td>c</td>
</tr>
<tr>
<td>4 tsize</td>
<td>s</td>
</tr>
<tr>
<td>5 tgrade</td>
<td>s</td>
</tr>
<tr>
<td>6 pnodes</td>
<td>s</td>
</tr>
<tr>
<td>7 progres</td>
<td>c</td>
</tr>
<tr>
<td>8 estrec</td>
<td>s</td>
</tr>
<tr>
<td>9 time</td>
<td>t</td>
</tr>
<tr>
<td>10 death</td>
<td>d</td>
</tr>
</tbody>
</table>

============= Constructed variables ==============

11 horTh.yes f 0.000 1.000

Total # cases w/ # missing

# cases miss. D ord. vals #X-var #N-var #F-var #S-var
686 0 0 0 0 0 6

# P-var # M-var # B-var # C-var # I-var # R-var
0 0 0 0 1 0 1

No weight variable in data file
Number of cases used for training: 533
Number of split variables: 7
Number of dummy variables created: 1

Constant fitted to cases with missing values in regressor variables
Predictive priority (Gi) using restricted mean event time
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE's for pruned tree: 0.2500

No nodewise interaction tests
Split values for N and S variables based on exhaustive search
Maximum number of split levels: 10
Minimum node sample size: 6
Minimum fraction of cases per treatment at each node: 0.073
Top-ranked variables and 1-df chi-squared values at root node

1 0.1169E+02 estrec
2 0.2062E+01 progres
3 0.1847E+01 tgrade
4 0.4400E+00 age
5 0.3773E+00 pnodes

Wei-Yin Loh

GUIDE manual
11.3 Censored response: restricted mean

Size and CV MSE and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Tnodes</th>
<th>Mean MSE</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median MSE</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>55</td>
<td>5.137E+05</td>
<td>2.795E+04</td>
<td>1.644E+04</td>
<td>5.090E+05</td>
<td>2.925E+04</td>
</tr>
<tr>
<td>2</td>
<td>54</td>
<td>5.137E+05</td>
<td>2.795E+04</td>
<td>1.644E+04</td>
<td>5.090E+05</td>
<td>2.925E+04</td>
</tr>
<tr>
<td></td>
<td>:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>5</td>
<td>5.035E+05</td>
<td>2.647E+04</td>
<td>1.048E+04</td>
<td>5.064E+05</td>
<td>1.206E+04</td>
</tr>
<tr>
<td>35+</td>
<td>2</td>
<td>4.463E+05</td>
<td>2.216E+04</td>
<td>1.042E+04</td>
<td>4.353E+05</td>
<td>2.040E+04</td>
</tr>
<tr>
<td>36**</td>
<td>1</td>
<td>4.338E+05</td>
<td>1.732E+04</td>
<td>6.012E+03</td>
<td>4.385E+05</td>
<td>7.335E+03</td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 1 terminal node
0-SE tree based on median is marked with + and has 2 terminal node
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
** tree same as ++ tree
** tree same as -- tree
++ tree same as -- tree
* tree same as ** tree
* tree same as ++ tree
* tree same as -- tree

Following tree is based on mean CV with naive SE estimate (**)
Node 1: Terminal node

Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>960.8</td>
<td>51.78</td>
<td>0.000</td>
<td>0.000</td>
<td>0.3591</td>
<td>1.000</td>
</tr>
<tr>
<td>horTh.yes</td>
<td>73.85</td>
<td>2.385</td>
<td>0.1744E-01</td>
<td>0.000</td>
<td>0.3591</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Time mean = 987.273

No truncation of predicted values

----------------------------------------

Observed and fitted values are stored in rest-gi.fit
LaTeX code for tree is in rest-gi.tex
R code is stored in rest-gi.r

Results for Gs method  The piecewise-constant Gs tree has one split, as shown below.

Restricted mean event time regression tree
Pruning by cross-validation
Data description file: cancerdsc.txt
Training sample file: cancerdata.txt
Missing value code: NA
Records in data file start on line 2
R variable present
6 N variables changed to S
Warning: model changed to linear in treatment
D variable is death
Piecewise linear model
Number of records in data file: 686
Length of longest entry in data file: 4
Treatment (R) variable is horTh with values "no" and "yes"
Number of dummy variables created: 1
Smallest uncensored time: 72.0000
Largest uncensored and censored time by horTh

<table>
<thead>
<tr>
<th>horTh</th>
<th>Uncensored</th>
<th>Censored</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;no&quot;</td>
<td>2456.0000</td>
<td>2563.0000</td>
</tr>
<tr>
<td>&quot;yes&quot;</td>
<td>2372.0000</td>
<td>2659.0000</td>
</tr>
</tbody>
</table>

Interval for restricted mean event time is from 0 to 1222.
Proportion of training sample for each level of horTh

| "no" | 0.6360 |
| "yes" | 0.3640 |

Summary information for training sample of size 533 (excluding observations with non-positive weight or missing values in d, e, t, r or z variables)
d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators),
11.3  Censored response: restricted mean  11  RANDOMIZED TREATMENTS

s=s-split-only numerical, n=split and fit numerical, f=fit-only numerical,
m=missing-value flag variable, p=periodic variable, w=weight

#Codes/  Levels/
Column  Name  Minimum  Maximum  Periods  #Missing
  1  horTh  r  2
  2  age  s  21.00  80.00
  3  menostat  c  2
  4  tsize  s  3.000  120.0
  5  tgrade  s  1.000  3.000
  6  pnodes  s  1.000  36.00
  7  progres  s  0.000  1490.
  8  estrec  s  0.000  1091.
  9  time  t  72.00  2659.
 10  death  d  0.000  1.000

================================== Constructed variables ====================
  11  horTh.yes  f  0.000  1.000

Total  #cases w/  #missing
  #cases  miss.  D  ord.  vals  #X-var  #N-var  #F-var  #S-var
  686  0  0  0  0  0  0  0
  #P-var  #M-var  #B-var  #C-var  #I-var  #R-var
  0  0  0  1  0  1

No weight variable in data file
Number of cases used for training: 533
Number of split variables: 7
Number of dummy variables created: 1

Constant fitted to cases with missing values in regressor variables
Prognostic priority (Gs) using restricted mean event time
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE’s for pruned tree: 0.2500

No nodewise interaction tests
Split values for N and S variables based on exhaustive search
Maximum number of split levels: 10
Minimum node sample size: 6
Minimum fraction of cases per treatment at each node: 0.073
Top-ranked variables and 1-df chi-squared values at root node
  1  0.4966E+02  pnodes
  2  0.3191E+02  progres
  3  0.2229E+02  estrec
  4  0.1276E+02  tgrade
  5  0.6795E+01  tsize
  6  0.4436E+01  age

Wei-Yin Loh  225  GUIDE manual
11.3  Censored response: restricted mean  11  RANDOMIZED TREATMENTS

7  0.1645E+00  menostat

Size and CV MSE and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Nodes</th>
<th>Mean MSE</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median MSE</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>58</td>
<td>4.781E+05</td>
<td>2.651E+04</td>
<td>2.824E+04</td>
<td>4.735E+05</td>
<td>3.191E+04</td>
</tr>
<tr>
<td>2</td>
<td>57</td>
<td>4.781E+05</td>
<td>2.651E+04</td>
<td>2.824E+04</td>
<td>4.735E+05</td>
<td>3.191E+04</td>
</tr>
<tr>
<td>:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>3</td>
<td>4.236E+05</td>
<td>2.187E+04</td>
<td>1.596E+04</td>
<td>4.273E+05</td>
<td>3.065E+04</td>
</tr>
<tr>
<td>39**</td>
<td>2</td>
<td>3.798E+05</td>
<td>1.852E+04</td>
<td>1.523E+04</td>
<td>3.804E+05</td>
<td>1.576E+04</td>
</tr>
<tr>
<td>40</td>
<td>1</td>
<td>4.338E+05</td>
<td>1.732E+04</td>
<td>6.012E+03</td>
<td>4.385E+05</td>
<td>7.335E+03</td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 2 terminal nodes
0-SE tree based on median is marked with + and has 2 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on median and bootstrap SE is marked with ++
* tree, ** tree, + tree, and ++ tree all the same

Following tree is based on mean CV with naive SE estimate (**)  

Structure of final tree. Each terminal node is marked with a T.

D-mean is weighted mean of death in the node
Cases fit give the number of cases used to fit node
MSE and R^2 are based on all cases in node

<table>
<thead>
<tr>
<th>Node</th>
<th>Total Cases</th>
<th>Matrix Node</th>
<th>Cases fit</th>
<th>Rank</th>
<th>D-mean</th>
<th>MSE</th>
<th>R^2</th>
<th>variable</th>
<th>Other variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>533</td>
<td>533</td>
<td>2</td>
<td>9.873E+02</td>
<td>1.519E+05</td>
<td>0.0106</td>
<td>pnodes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2T</td>
<td>332</td>
<td>332</td>
<td>2</td>
<td>1.073E+02</td>
<td>1.048E+05</td>
<td>0.0129</td>
<td>estrec</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3T</td>
<td>201</td>
<td>201</td>
<td>2</td>
<td>8.312E+02</td>
<td>1.842E+05</td>
<td>0.0174</td>
<td>progrec</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Number of terminal nodes of final tree: 2
Total number of nodes of final tree: 3
Second best split variable (based on curvature test) at root node is progrec

Regression tree:

Node 1: pnodes <= 4.500000
   Node 2: terminal
Node 1: pnodes > 4.500000 or NA
   Node 3: terminal

***************************************************************************************

WARNING: p-values below not adjusted for split search. For a bootstrap solution see:

Wei-Yin Loh  226  GUIDE manual
11.3 Censored response: restricted mean


Node 1: Intermediate node
A case goes into Node 2 if pnodes <= 4.500000

Coefficients of least squares regression function:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>960.8</td>
<td>51.78</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>horTh.yes</td>
<td>73.85</td>
<td>2.385</td>
<td>0.1744E-01</td>
<td>0.000</td>
<td>0.3591</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Time mean = 987.273
No truncation of predicted values
----------------------------------------

Node 2: Terminal node

Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1050.</td>
<td>55.19</td>
<td>0.2220E-15</td>
<td>0.000</td>
<td>0.3483</td>
<td>1.000</td>
</tr>
<tr>
<td>horTh.yes</td>
<td>66.83</td>
<td>2.074</td>
<td>0.3884E-01</td>
<td>0.000</td>
<td>0.3483</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Time mean = 1072.91
No truncation of predicted values
----------------------------------

Node 3: Terminal node

Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>790.8</td>
<td>22.68</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>horTh.yes</td>
<td>106.5</td>
<td>1.879</td>
<td>0.6164E-01</td>
<td>0.000</td>
<td>0.3786</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Time mean = 831.171
No truncation of predicted values
----------------------------------

Observed and fitted values are stored in rest-gs.fit
LaTeX code for tree is in rest-gs.tex
R code is stored in rest-gs.r

11.3.2 With linear prognostic control

A trivial tree is obtained for both the Gi and Gs methods if a linear regressor is included in each node.
12 Nonrandomized treatments: RHC data

A classification tree was built in Section 4 to predict the occurrence of right heart catheterization (RHC), which is a treatment used to treat critically ill patients with heart problems. GUIDE can fit a tree model to find subgroups where the treatment (represented by variable \textit{swang1}) is beneficial or not for survival. This is done by specifying the treatment variable as “r” and the event variable \textit{death} (1=die, 0=not die) as “d” in the description file \texttt{rhcdsc3.txt} below.

\begin{verbatim}
rhcdsc3.txt
NA
2
1 X x
2 cat1 c
3 cat2 c
4 ca c
5 sadmdte x
6 dschdte x
7 dthdte x
8 lstctdte x
9 death d
10 cardiohx c
11 chfhx c
12 dementhx c
13 psychhx c
14 chrpulhx c
15 renalhx c
16 liverhx c
17 gibledhx c
18 malighx c
19 immunhx c
20 transhx c
21 amihx c
22 age n
23 sex c
24 edu n
25 surv2md1 n
26 das2d3pc n
27 t3d30 x
28 dth30 x
29 aps1 n
30 scoma1 n
31 meanbp1 n
32 wblc1 n
33 hrt1 n
34 respi n
\end{verbatim}
12.1 Proportional hazards

GUIDE can fit models with the Gi or Gs options. The Gi option is designed to be sensitive to detect predictive variables (variables that have interactions with the treatment variable) while Gs option is equally sensitive to such variables as well as prognostic variables (those that have an effect on the outcome irrespective of the treatment). See Loh et al. (2015) for details.

12.1.1 Gi option

Gi input file creation

0. Read the warranty disclaimer

Wei-Yin Loh 229 GUIDE manual
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: surv-gi.in
   Input 1 for model fitting, 2 for importance or DIF scoring,
   3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: surv-gi.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1): 2
   Choose type of regression model:
     1=linear, 2=quantile, 3=Poisson, 4=censored response,
     5=multiresponse or itemresponse, 6=longitudinal data (with T variables),
     7=binary logistic regression.
   Input choice ([1:7], <cr>=1): 4
   Input 1 for proportional hazards, 2 for restricted mean event time ([1:2], <cr>=1): 4
   Choose complexity of model to use at each node:
   Choose 1 for multiple regression (recommended for prediction)
   Choose 2 for best simple linear in one N or F variable
   Choose 3 for constant fit (recommended for interpretability or if there is an R variable)
   1: multiple linear, 2: best simple linear, 3: constant ([1:3], <cr>=3):
   Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
   Input name of data description file (max 100 characters);
   enclose with matching quotes if it has spaces: rhcdsc3.txt
Reading data description file ...
Training sample file: rhcdata.txt
Missing value code: NA
Records in data file start on line 2
R variable present
23 N variables changed to S
Warning: model changed to linear in treatment
D variable is death
Reading data file ...
Number of records in data file: 5735
Length of longest entry in data file: 19
Checking for missing values ...
Finished checking
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Assigning integer codes to values of 31 categorical variables
Finished assigning codes to 10 categorical variables
Finished assigning codes to 20 categorical variables
Finished assigning codes to 30 categorical variables
Treatment (R) variable is swang1 with values "NoRHC" and "RHC"
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any...
Finished processing 5000 of 5735 observations
Data checks complete
Smallest uncensored survtime: 2.0000
Number of complete cases excluding censored T < smallest uncensored T: 5735
Number of cases used to compute baseline hazard: 5735
Number of cases with D=1 and T >= smallest uncensored: 3722
GUIDE will try to create the variables in the description file.
If it is unsuccessful, please create the columns yourself...
Number of dummy variables created: 1
Choose a subgroup identification method:
1 = Prognostic priority (Gs)
2 = Predictive priority (Gi)
Input your choice: ([1:2], <cr>=2):
Creating dummy variables ...
Creating missing value indicators ...
Rereading data ...
Largest uncensored and censored survtime by swang1
"NoRHC" 1867.0000 1243.0000
"RHC" 1943.0000 1351.0000
Proportion of training sample for each level of swang1
"NoRHC" 0.6192
"RHC" 0.3808

Total #cases w/ #missing
#cases miss. D ord. vals #X-var #N-var #F-var #S-var
5735 0 5157 8 0 0 23
#P-var #M-var #B-var #C-var #I-var #R-var
0 0 0 30 0 1
Survival time variable in column: 64
Event indicator variable in column: 9
Proportion uncensored among nonmissing T and D variables: .649
Number of cases used for training: 5735
Number of split variables: 53
Number of dummy variables created: 1
Number of cases excluded due to 0 W or missing D, T or R variables: 0
Finished reading data file
Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):
Input file name to store LaTeX code (use .tex as suffix): surv-gi.tex
You can store the variables and/or values used to split and fit in a file
Choose 1 to skip this step, 2 to store split and fit variables,
3 to store split variables and their values
Input your choice ([1:3], <cr>=1):
Input 2 to save fitted values and node IDs, 1 otherwise ([1:2], <cr>=2):
Input name of file to store node ID and fitted value of each case: surv-gi.fit
Input 2 to write R function for predicting new cases, 1 otherwise ([1:2], <cr>=2):
Input file name: surv-gi.r
12.1 Proportional hazards

Input rank of top variable to split root node ([1:55], <cr>=1):
Input file is created!
Run GUIDE with the command: guide < surv-gi.in

Contents of surv-gi.out

Regression tree for censored response
Pruning by cross-validation
Data description file: rhcdsc3.txt
Training sample file: rhcdata.txt
Missing value code: NA
Records in data file start on line 2
R variable present
23 N variables changed to S
Warning: model changed to linear in treatment
D variable is death
Piecewise linear model
Number of records in data file: 5735
Length of longest entry in data file: 19
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Treatment (R) variable is swang1 with values "NoRHC" and "RHC"
Number of complete cases excluding censored T < smallest uncensored T: 5735
Number of cases used to compute baseline hazard: 5735
Number of cases with D=1 and T >= smallest uncensored: 3722
Number of dummy variables created: 1
Smallest uncensored survtime: 2.0000
Largest uncensored and censored survtime by swang1

swang1 Uncensored Censored
"NoRHC" 1867.0000 1243.0000
"RHC" 1943.0000 1351.0000
Proportion of training sample for each level of swang1
"NoRHC" 0.6192
"RHC" 0.3808

Summary information for training sample of size 5735

#Codes/
Levels/
Column Name Minimum Maximum Periods #Missing

Wei-Yin Loh 232 GUIDE manual
## 12.1 Proportional hazards

### NONRANDOMIZED TREATMENTS: RHC DATA

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cat1</td>
<td>c</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cat2</td>
<td>c</td>
<td>6</td>
<td>4535</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ca</td>
<td>c</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>death</td>
<td>d</td>
<td>0.000</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cardiohx</td>
<td>c</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>adld3p</td>
<td>s</td>
<td>0.000</td>
<td>7.000</td>
<td>4296</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>urin1</td>
<td>s</td>
<td>0.000</td>
<td>9000.0</td>
<td>3028</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>race</td>
<td>c</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>income</td>
<td>c</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>survtime</td>
<td>t</td>
<td>2.000</td>
<td>1943</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

====== Constructed variables ======

|   |   |   |   |   |
|---|---|---|---|
|   | lnbsehaz | z | -3.818 | 2.038 |   |   |
|   | swang1.RHC | f | 0.000 | 1.000 |   |   |

---

<table>
<thead>
<tr>
<th>Total cases w/ #missing</th>
<th>#cases</th>
<th>miss. D</th>
<th>ord. vals</th>
<th>#X-var</th>
<th>#N-var</th>
<th>#F-var</th>
<th>#S-var</th>
</tr>
</thead>
<tbody>
<tr>
<td>5735</td>
<td>5157</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#P-var</th>
<th>#M-var</th>
<th>#B-var</th>
<th>#C-var</th>
<th>#I-var</th>
<th>#R-var</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>30</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Survival time variable in column: 64
Event indicator variable in column: 9
Proportion uncensored among nonmissing T and D variables: 0.649
Number of cases used for training: 5735
Number of split variables: 53
Number of dummy variables created: 1
Number of cases excluded due to 0 W or missing D, T or R variables: 0

Predictive priority (Gi)
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE’s for pruned tree: 0.2500

No nodewise interaction tests
Fraction of cases used for splitting each node: 1.0000
Maximum number of split levels: 15
Minimum node sample size: 7
Minimum fraction of cases per treatment at each node: 0.076
Number of iterations for fitting: 20
Top-ranked variables and 1-df chi-squared values at root node

|   |   |   |   |   |
|---|---|---|---|
| 1 | 0.1323E+02 | ph1 |   |
| 2 | 0.1018E+02 | resp1 |   |
| 3 | 0.8324E+01 | cat2 |   |
| 4 | 0.7453E+01 | pot1 |   |
|   |   |   |   |
| 35 | 0.1497E-01 | sod1 |   |

---

Wei-Yin Loh 233 GUIDE manual
12.1 Proportional hazards  

NONRANDOMIZED TREATMENTS: RHC DATA

Size and CV Loss and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Nodes</th>
<th>Mean Loss</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median Loss</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>499</td>
<td>2.105E+00</td>
<td>6.751E-02</td>
<td>5.520E-02</td>
<td>2.061E+00</td>
<td>5.624E-02</td>
</tr>
<tr>
<td>2</td>
<td>498</td>
<td>2.105E+00</td>
<td>6.751E-02</td>
<td>5.520E-02</td>
<td>2.061E+00</td>
<td>5.624E-02</td>
</tr>
<tr>
<td>321</td>
<td>14</td>
<td>1.323E+00</td>
<td>1.610E-02</td>
<td>6.606E-03</td>
<td>1.334E+00</td>
<td>1.298E-02</td>
</tr>
<tr>
<td>322**</td>
<td>5</td>
<td>1.322E+00</td>
<td>1.586E-02</td>
<td>7.111E-03</td>
<td>1.331E+00</td>
<td>1.190E-02</td>
</tr>
<tr>
<td>323</td>
<td>1</td>
<td>1.367E+00</td>
<td>1.526E-02</td>
<td>6.317E-03</td>
<td>1.358E+00</td>
<td>9.980E-03</td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 5 terminal nodes
0-SE tree based on median is marked with + and has 5 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on median and bootstrap SE is marked with ++
* tree, ** tree, + tree, and ++ tree all the same

Following tree is based on mean CV with naive SE estimate (**)

Structure of final tree. Each terminal node is marked with a T.

Cases fit give the number of cases used to fit node
Deviance is mean residual deviance for all cases in node

<table>
<thead>
<tr>
<th>Node</th>
<th>Total Cases</th>
<th>Matrix Median</th>
<th>Node Split</th>
</tr>
</thead>
<tbody>
<tr>
<td>label</td>
<td>cases fit rank</td>
<td>surve time deviance variable</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5735 5735</td>
<td>1 1.920E+02</td>
<td>1.367E+00</td>
</tr>
<tr>
<td>2</td>
<td>1411 1411</td>
<td>1 1.150E+02</td>
<td>1.454E+00</td>
</tr>
<tr>
<td>4T</td>
<td>1307 1307</td>
<td>1 1.570E+02</td>
<td>1.416E+00</td>
</tr>
<tr>
<td>5T</td>
<td>104 104</td>
<td>1 1.400E+01</td>
<td>1.636E+00</td>
</tr>
<tr>
<td>3</td>
<td>4324 4324</td>
<td>1 2.070E+02</td>
<td>1.334E+00</td>
</tr>
<tr>
<td>6</td>
<td>3341 3341</td>
<td>1 2.200E+02</td>
<td>1.333E+00</td>
</tr>
<tr>
<td>12T</td>
<td>687 687</td>
<td>1 6.900E+01</td>
<td>1.531E+00</td>
</tr>
<tr>
<td>13T</td>
<td>2654 2654</td>
<td>1 2.390E+02</td>
<td>1.265E+00</td>
</tr>
<tr>
<td>7T</td>
<td>983 983</td>
<td>1 1.640E+02</td>
<td>1.319E+00</td>
</tr>
</tbody>
</table>

Number of terminal nodes of final tree: 5
Total number of nodes of final tree: 9
Second best split variable (based on curvature test) at root node is resp1

Regression tree:
For categorical variable splits, values not in training data go to the right

Node 1: ph1 <= 7.3344730
Node 2: cat2 = "MDSF w/Sepsis", "NA"
Node 4: Median survival time = 157.00000

Wei-Yin Loh 234 GUIDE manual
12.1 Proportional hazards

| Node 2: cat2 /= "MOSF w/Sepsis", "NA" |
| Node 5: Median survival time = 14.000000 |
| Node 1: ph1 > 7.3344730 or NA |
| Node 3: resp1 <= 38.500000 or NA |
| Node 6: paco21 <= 29.498050 |
| Node 12: Median survival time = 69.000000 |
| Node 6: paco21 > 29.498050 or NA |
| Node 13: Median survival time = 239.000000 |
| Node 3: resp1 > 38.500000 |
| Node 7: Median survival time = 164.000000 |

***************************************************************
Predictor means below are means of cases with no missing values.
Regression coefficients are computed from the complete cases.

**WARNING:** p-values below not adjusted for split search. For a bootstrap solution see:


<p>| Node 1: Intermediate node |
| A case goes into Node 2 if ph1 &lt;= 7.3344730 |
| ph1 mean = 7.3884135 |
| Coefficients of log-relative hazard function (relative to baseline hazard): |</p>
<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>swang1.RHC</td>
<td>0.1504</td>
<td>4.494</td>
<td>0.7131E-05</td>
<td>0.000</td>
<td>0.3808</td>
<td>1.000</td>
</tr>
</tbody>
</table>

---------------------------------------------------------------------

| Node 2: Intermediate node |
| A case goes into Node 4 if cat2 = "MOSF w/Sepsis", "NA" |
| cat2 mode = "NA" |

---------------------------------------------------------------------

<p>| Node 4: Terminal node |
| Coefficients of log-relative hazard function (relative to baseline hazard): |</p>
<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.6181E-01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>swang1.RHC</td>
<td>0.4067</td>
<td>6.034</td>
<td>0.2086E-08</td>
<td>0.000</td>
<td>0.4499</td>
<td>1.000</td>
</tr>
</tbody>
</table>

---------------------------------------------------------------------

| Node 5: Terminal node |
| Coefficients of log-relative hazard function (relative to baseline hazard): |
### 12.1 Proportional hazards

#### NONRANDOMIZED TREATMENTS: RHC DATA

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.8005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>swang1.RHC</td>
<td>-0.3295</td>
<td>-1.558</td>
<td>0.1223</td>
<td>0.000</td>
<td>0.3558</td>
<td>1.000</td>
</tr>
</tbody>
</table>

---

**Node 3:** Intermediate node

A case goes into Node 6 if \( \text{resp1} \leq 38.500000 \) or \( \text{NA} \)

\( \text{resp1 mean} = 28.418652 \)

---

**Node 6:** Intermediate node

A case goes into Node 12 if \( \text{paco21} \leq 29.498050 \)

\( \text{paco21 mean} = 36.054906 \)

---

**Node 12:** Terminal node

Coefficients of log-relative hazard function (relative to baseline hazard):

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.3006</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>swang1.RHC</td>
<td>-0.3237E-01</td>
<td>-0.3424</td>
<td>0.7322</td>
<td>0.000</td>
<td>0.3916</td>
<td>1.000</td>
</tr>
</tbody>
</table>

---

**Node 13:** Terminal node

Coefficients of log-relative hazard function (relative to baseline hazard):

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.7105E-01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>swang1.RHC</td>
<td>0.5937E-02</td>
<td>0.1159</td>
<td>0.9078</td>
<td>0.000</td>
<td>0.3632</td>
<td>1.000</td>
</tr>
</tbody>
</table>

---

**Node 7:** Terminal node

Coefficients of log-relative hazard function (relative to baseline hazard):

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.1150E-01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>swang1.RHC</td>
<td>0.3555</td>
<td>4.329</td>
<td>0.1651E-04</td>
<td>0.000</td>
<td>0.3316</td>
<td>1.000</td>
</tr>
</tbody>
</table>

---

Observed and fitted values are stored in surv.gi.fit

LaTeX code for tree is in surv-gi.tex

R code is stored in surv-gi.r

---

Figure 28 shows the tree and Figure 29 shows the estimated survival curves in its terminal nodes. The R code for making the plots is given below.

```r
library(survival)
z0 <- read.table("rhcdata.txt",header=TRUE)
par(mar=c(3,4,3,1),mfrow=c(2,3),cex=1)
leg.txt <- c("NoRHC","RHC"); leg.col <- c("blue","red"); leg.lty <- 2:1
xr <- range(z0$survtime)
zg <- read.table("surv-gi.fit",header=TRUE)
nodes <- zg$node
uniq.gp <- unique(sort(nodes))
```

Wei-Yin Loh 236  GUIDE manual
Figure 28: GUIDE v.41.1 0.250-SE proportional hazards regression tree using Gi option for `survtime` and event indicator `death` without adjustment for linear prognostic effects. At each split, an observation goes to the left branch if and only if the condition is satisfied. Symbol ‘≤∗’ stands for ‘≤ or missing’. \( S_1 = \{\text{MOSF w/Sepsis, NA}\} \). Treatment `swang1` hazard ratio of level `RHC` to level `NoRHC` beside nodes. Sample size (in italics), median survival time, and proportion of `swang1 = RHC` printed below nodes. Terminal nodes with treatment hazard ratio above and below 1.162 (ratio at root node) are colored orange and skyblue respectively. Second best split variable at root node is `resp1`.

Wei-Yin Loh 237 GUIDE manual
12.1 Proportional hazards

Figure 29: Survival curves for RHC data in nodes of Figure 28
Following are the top 3 lines of the file `surv-gi.fit`

<table>
<thead>
<tr>
<th>train</th>
<th>node</th>
<th>observed</th>
<th>event</th>
<th>logbasecumhaz</th>
<th>survivalprob</th>
<th>mediansurvtme</th>
<th>swang1.RHC</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>13</td>
<td>240.000</td>
<td>n</td>
<td>-0.269165</td>
<td>0.490850</td>
<td>239.000</td>
<td>0.593672E-002</td>
</tr>
<tr>
<td>y</td>
<td>4</td>
<td>45.0000</td>
<td>y</td>
<td>-0.757608</td>
<td>0.515901</td>
<td>157.000</td>
<td>0.406690</td>
</tr>
<tr>
<td>y</td>
<td>7</td>
<td>317.000</td>
<td>n</td>
<td>-0.633003E-001</td>
<td>0.266047</td>
<td>164.000</td>
<td>0.355517</td>
</tr>
</tbody>
</table>

The column definitions are

**train:** y if the observation is used for model fitting, n if not.

**node:** terminal node label of observation.

**observed:** observed survival time $t$.

**event:** y if uncensored (death), n if censored.

**logbasecumhaz:** log of the estimated baseline cumulative hazard function $\log \Lambda_0(t) = \log \int_0^t \lambda_0(u) \, du$ at observed time $t$.

**survivalprob:** probability that the subject survives up to observed time $t$. For the first subject, this is

$$\exp\left\{-\Lambda_0(t) \exp(\beta'x)\right\} = \exp\{-\exp(\beta_0 + \logbasecumhaz)\}$$

$$= \exp(-\exp(-0.242135921383 - 0.3029494))$$

$$= 0.5600147$$
where $t = 240$ and $\beta_0 = -0.242135921383$ is the constant term in the node
(surv-gs.r gives $\beta_0$ to higher precision than surv-gs.out).

\textbf{mediansurvtime}: median survival time among observations in node estimated from
Kaplan-Meier survival function. A trailing plus (+) sign indicates estimate is
censored.

\textbf{swang1.RHC}: estimated treatment effect $\beta_1$ for level RHC of \textbf{swang1}.

### 12.2 Restricted mean

GUIDE can also construct a tree model such that a restricted mean event time
(Chen and Tsiatis, 2001; Tian et al., 2014) is fitted in each node of the tree.

#### 12.2.1 Gi option

\textbf{Gi input file creation}

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: rest-gi.in
   Input 1 for model fitting, 2 for importance or DIF scoring,
   3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: rest-gi.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1): 2
   Choose type of regression model:
   1=linear, 2=quantile, 3=Poisson, 4=censored response,
   5=multiplesresponse or itemresponse, 6=longitudinal data (with T variables),
   7=binary logistic regression.
   Input choice ([1:7], <cr>=1): 4
   Input 1 for proportional hazards, 2 for restricted mean event time ([1:2], <cr>=1): 2
   Choose complexity of model to use at each node:
   Choose 0 for stepwise linear regression (recommended for prediction)
   Choose 1 for multiple regression
   Choose 2 for best simple polynomial in one N or F variable
   Choose 3 for constant fit (recommended for interpretability or if there is an R variable)
   0: stepwise linear, 1: multiple linear, 2: best simple polynomial, 3: constant,
   4: best simple stepwise ANCOVA ([0:4], <cr>=3):
   Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
   Input name of data description file (max 100 characters);
   enclose with matching quotes if it has spaces: rhcdsc3.txt
12.2 Restricted mean

Reading data description file ...
Training sample file: rhcdata.txt
Missing value code: NA
Records in data file start on line 2
R variable present
23 N variables changed to S
Warning: model changed to linear in treatment
D variable is death
Reading data file ...
Number of records in data file: 5735
Length of longest entry in data file: 19
Checking for missing values ...
Finished checking
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Assigning integer codes to values of 31 categorical variables
Finished assigning codes to 10 categorical variables
Finished assigning codes to 20 categorical variables
Finished assigning codes to 30 categorical variables
Treatment (R) variable is swang1 with values "NoRHC" and "RHC"
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...
Finished processing 5000 of 5735 observations
Data checks complete
GUIDE will try to create the variables in the description file.
If it is unsuccessful, please create the columns yourself...
Number of dummy variables created: 1
Choose a subgroup identification method:
1 = Prognostic priority (Gs)
2 = Predictive priority (Gi)
Input your choice: ([1:2], <cr>=2):
Creating dummy variables ...
Creating missing value indicators ...
Rereading data ...
Largest uncensored and censored survtime by swang1
"NoRHC"  1867.0000  1243.0000
"RHC"  1943.0000  1351.0000
Smallest observed uncensored time is 2.0000
Largest observed censored or uncensored time is 1943.0000
Input restriction on event time ([2.00:1943.00], <cr>=622.00):
Proportion of training sample for each level of swang1
"NoRHC"  0.5993
"RHC"  0.4007
Total #cases w/ #missing
# Restricted mean event time regression tree

Pruning by cross-validation

Data description file: rhcdsc3.txt

Training sample file: rhcdata.txt

Missing value code: NA

Records in data file start on line 2

R variable present

23 N variables changed to S

Warning: model changed to linear in treatment

D variable is death

Piecewise linear model

Number of records in data file: 5735

Length of longest entry in data file: 19

Missing values found among categorical variables

Separate categories will be created for missing categorical variables

Missing values found among non-categorical variables

Treatment (R) variable is swang1 with values "NoRHC" and "RHC"

Number of dummy variables created: 1

Smallest uncensored survtime: 2.0000
Largest uncensored and censored survival time by swang1

<table>
<thead>
<tr>
<th></th>
<th>Uncensored</th>
<th>Censored</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;NoRHC&quot;</td>
<td>1867.0000</td>
<td>1243.0000</td>
</tr>
<tr>
<td>&quot;RHC&quot;</td>
<td>1943.0000</td>
<td>1351.0000</td>
</tr>
</tbody>
</table>

Interval for restricted mean event time is from 0 to 622.

Proportion of training sample for each level of swang1

- "NoRHC": 0.5993
- "RHC": 0.4007

Summary information for training sample of size 3763 (excluding observations with non-positive weight or missing values in d, e, t, r or z variables)

- d=dependent, b=split and fit cat variable using indicator variables,
- c=split-only categorical, i=fit-only categorical (via indicators),
- s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
- m=missing-value flag variable, w=weight

<table>
<thead>
<tr>
<th>Code/Levels/</th>
<th>#Codes/Levels/</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 cat1 c</td>
<td>Minimum</td>
<td>Maximum</td>
</tr>
<tr>
<td>3 cat2 c</td>
<td>6</td>
<td>2836</td>
</tr>
<tr>
<td>4 ca c</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>9 death d</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>61 race c</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>62 income c</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>64 survtime t</td>
<td>2.000</td>
<td>1943.</td>
</tr>
</tbody>
</table>

=================== Constructed variables ===================

<table>
<thead>
<tr>
<th></th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>65 swang1.RHC f</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Total #cases w/ #missing

<table>
<thead>
<tr>
<th>#cases</th>
<th>miss. D</th>
<th>ord. vals</th>
<th>#X-var</th>
<th>#N-var</th>
<th>#F-var</th>
<th>#S-var</th>
</tr>
</thead>
<tbody>
<tr>
<td>5735</td>
<td>5157</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td></td>
<td>#P-var</td>
<td>#M-var</td>
<td>#B-var</td>
<td>#C-var</td>
<td>#I-var</td>
<td>#R-var</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

No weight variable in data file

Number of cases used for training: 3763

Number of split variables: 53

Number of dummy variables created: 1

Number of cases excluded due to 0 W or missing D or R variables: 1972

Predictive priority (Gi) using restricted mean event time

Pruning by v-fold cross-validation, with v = 10

Selected tree is based on mean of CV estimates

Number of SE's for pruned tree: 0.2500

No nodewise interaction tests
Split values for N and S variables based on exhaustive search
Maximum number of split levels: 13
Minimum node sample size: 7
Minimum fraction of cases per treatment at each node: 0.080
Top-ranked variables and 1-df chi-squared values at root node

1 0.9407E+01 scoma1
2 0.7887E+01 ph1
3 0.7551E+01 pafi1

Size and CV MSE and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Nodes</th>
<th>Mean MSE</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median MSE</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>325</td>
<td>1.644E+05</td>
<td>5.598E+03</td>
<td>3.914E+03</td>
<td>1.652E+05</td>
<td>6.405E+03</td>
</tr>
<tr>
<td>2</td>
<td>324</td>
<td>1.644E+05</td>
<td>5.598E+03</td>
<td>3.914E+03</td>
<td>1.652E+05</td>
<td>6.405E+03</td>
</tr>
<tr>
<td>3</td>
<td>323</td>
<td>1.644E+05</td>
<td>5.598E+03</td>
<td>3.914E+03</td>
<td>1.652E+05</td>
<td>6.403E+03</td>
</tr>
<tr>
<td>217</td>
<td>3</td>
<td>1.295E+05</td>
<td>4.444E+03</td>
<td>4.786E+03</td>
<td>1.294E+05</td>
<td>6.909E+03</td>
</tr>
<tr>
<td>218**</td>
<td>2</td>
<td>1.157E+05</td>
<td>3.411E+03</td>
<td>2.378E+03</td>
<td>1.141E+05</td>
<td>3.229E+03</td>
</tr>
<tr>
<td>219</td>
<td>1</td>
<td>1.198E+05</td>
<td>3.143E+03</td>
<td>9.972E+02</td>
<td>1.190E+05</td>
<td>1.421E+03</td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 2 terminal nodes
0-SE tree based on median is marked with + and has 2 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on median and bootstrap SE is marked with ++
* tree, ** tree, + tree, and ++ tree all the same

Following tree is based on mean CV with naive SE estimate (**)

Structure of final tree. Each terminal node is marked with a T.

D-mean is weighted mean of death in the node
Cases fit give the number of cases used to fit node
MSE and R^2 are based on all cases in node

<table>
<thead>
<tr>
<th>Node</th>
<th>Total cases</th>
<th>Cases at root node</th>
<th>Other variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>label</td>
<td>fit</td>
<td>rank</td>
<td>D-mean</td>
</tr>
<tr>
<td>1</td>
<td>3763</td>
<td>3763</td>
<td>2 2.583E+02 9.489E+04 0.0043 scoma1</td>
</tr>
<tr>
<td>2T</td>
<td>3124</td>
<td>3124</td>
<td>2 2.781E+02 9.938E+04 0.0075 pafi1</td>
</tr>
<tr>
<td>3T</td>
<td>639</td>
<td>639</td>
<td>2 1.333E+02 4.975E+04 0.0016 sod1</td>
</tr>
</tbody>
</table>

Number of terminal nodes of final tree: 2
Total number of nodes of final tree: 3
Second best split variable (based on curvature test) at root node is ph1
### Regression tree:

Node 1: `scoma1 <= 49.500000`
- Node 2: terminal

Node 1: `scoma1 > 49.500000` or NA
- Node 3: terminal

Predictor means below are means of cases with no missing values. Regression coefficients are computed from the complete cases.

**WARNING:** p-values below not adjusted for split search. For a bootstrap solution see:


#### Node 1: Intermediate node

A case goes into Node 2 if `scoma1 <= 49.500000`

**scoma1 mean = 20.462797**

Coefficients of least squares regression function:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>271.2</td>
<td>52.27</td>
<td>0.000</td>
<td>0.000</td>
<td>0.3808</td>
<td>1.000</td>
</tr>
<tr>
<td><code>swang1.RHC</code></td>
<td>-33.80</td>
<td>-4.020</td>
<td>0.5926E-04</td>
<td>0.000</td>
<td>0.3949</td>
<td>1.000</td>
</tr>
</tbody>
</table>

**survtime mean = 258.284**

No truncation of predicted values

#### Node 2: Terminal node

Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>295.7</td>
<td>51.17</td>
<td>0.000</td>
<td>0.000</td>
<td>0.3949</td>
<td>1.000</td>
</tr>
<tr>
<td><code>swang1.RHC</code></td>
<td>-44.75</td>
<td>-4.866</td>
<td>0.1195E-05</td>
<td>0.000</td>
<td>0.2916</td>
<td>1.000</td>
</tr>
</tbody>
</table>

**survtime mean = 278.051**

No truncation of predicted values

#### Node 3: Terminal node

Coefficients of least squares regression functions:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>138.4</td>
<td>14.56</td>
<td>0.000</td>
<td>0.000</td>
<td>0.2916</td>
<td>1.000</td>
</tr>
<tr>
<td><code>swang1.RHC</code></td>
<td>-17.66</td>
<td>-1.003</td>
<td>0.3161</td>
<td>0.000</td>
<td>0.2916</td>
<td>1.000</td>
</tr>
</tbody>
</table>

**survtime mean = 133.272**

Wei-Yin Loh 245 GUIDE manual
Figure 30: GUIDE v.41.1 0.250-SE regression tree using Gi option for mean `survtime` restricted to less than 622.00 without adjustment for linear prognostic effects. At each split, an observation goes to the left branch if and only if the condition is satisfied. Sample size (in \textit{italics}) printed below nodes. Treatment `swang1` effects (relative to reference level `NoRHC`) of levels `RHC` (relative to `NoRHC`) beside nodes. Terminal nodes with treatment effect above and below -33.80 (effect at root node) are colored orange and skyblue respectively. Second best split variable at root node is `ph1`.

No truncation of predicted values

\begin{itemize}
\item Number of times Li-Martin approximation used = 394
\item Observed and fitted values are stored in \texttt{rest-gi.fit}
\item \LaTeX{} code for tree is in \texttt{rest-gi.tex}
\item R code is stored in \texttt{rest-gi.r}
\end{itemize}

Figure 30 shows the Gi restricted mean event time tree.

13 Multiresponse: NMES data

GUIDE has two options for fitting a piecewise-constant regression model to predict two or more dependent variables simultaneously (Loh and Zheng, 2013). The first (named \texttt{multiresponse} or option 5 in the input file) requires the number of dependent variables to be the same for each observation. Observations with missing values in one or more dependent variables are excluded. The second (named \texttt{longitudinal data} (\textit{with T variables}) or option 6 in the input file) requires each dependent variable to be associated with an observation time variable. It fits a model to all observations, including those with missing values in some dependent variables. The observation times are not required to be the same for all subjects, i.e., they may vary from subject to subject, but observations with missing times are excluded from model fitting. We demonstrate the first option in this section. The second option is
Table 10: Definitions of variables in NMES data

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ofp</td>
<td>number of physician office visits</td>
</tr>
<tr>
<td>ofnp</td>
<td>number of nonphysician office visits</td>
</tr>
<tr>
<td>opp</td>
<td>number of physician outpatient visits</td>
</tr>
<tr>
<td>opnp</td>
<td>number of nonphysician outpatient visits</td>
</tr>
<tr>
<td>emer</td>
<td>number of emergency room visits</td>
</tr>
<tr>
<td>hosp</td>
<td>number of hospitalizations</td>
</tr>
<tr>
<td>health</td>
<td>self-perceived health (poor, average, or excellent)</td>
</tr>
<tr>
<td>numchron</td>
<td>number of chronic conditions</td>
</tr>
<tr>
<td>adldiff</td>
<td>has condition that limits daily living (no, yes)</td>
</tr>
<tr>
<td>region</td>
<td>region of U.S. (midwest, northeast, west, other)</td>
</tr>
<tr>
<td>age</td>
<td>age in years</td>
</tr>
<tr>
<td>black</td>
<td>African American (no, yes)</td>
</tr>
<tr>
<td>gender</td>
<td>sex (female, male)</td>
</tr>
<tr>
<td>married</td>
<td>married (no, yes)</td>
</tr>
<tr>
<td>school</td>
<td>number of years of education</td>
</tr>
<tr>
<td>faminc</td>
<td>family income in $10,000</td>
</tr>
<tr>
<td>employed</td>
<td>employed (no, yes)</td>
</tr>
<tr>
<td>privins</td>
<td>covered by private insurance (no, yes)</td>
</tr>
<tr>
<td>medicaid</td>
<td>covered by Medicaid (no, yes)</td>
</tr>
</tbody>
</table>

The data file nmes.txt contains observations on 4406 subjects from a National Medical Expenditure Survey (NMES) conducted in 1987 and 1988. Table 10 gives the names of the variables and their definitions. The data were previously analyzed in Deb and Trivedi (1997), Cameron and Trivedi (1998, chap. 6), and Zeileis (2006). Here we construct a regression tree to predict the outcomes for the first 6 variables (ofp, ofnp, opp, opnp, emer, and hosp). The contents of the description file nmes.dsc follow.

nmes.txt
NA
1
1 ofp d
2 ofnp d
3 opp d
4 opnp d
5 emer d
6 hosp d
13.1 Input file creation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: mult.in
   Input 1 for model fitting, 2 for importance or DIF scoring,
   3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: mult.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1): 2
   Choose type of regression model:
   1=linear, 2=quantile, 3=Poisson, 4=censored response,
   5=multiproresponse or itemresponse, 6=longitudinal data (with T variables),
   7=binary logistic regression.
   Input choice ([1:7], <cr>=1): 5
   Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
   Input name of data description file (max 100 characters);
   enclose with matching quotes if it has spaces: nmes.dsc
   Reading data description file ...
   Training sample file: nmes.txt
   Missing value code: NA
   Records in data file start on line 1
   4 N variables changed to S
   Number of D variables: 6
   D variables are:
   ofp
   ofnp
   opp

Wei-Yin Loh 248 GUIDE manual
Multivariate or univariate split variable selection:
Choose multivariate if there is an order among the D variables;
choose univariate otherwise or if item response
Input 1 for multivariate, 2 for univariate ([1:2], <cr>=1): 2
D variables can be normalized to have unit variance,
e.g., if they have different scales or units
Input 1 to normalize D variables, 2 for no normalization ([1:2], <cr>=1):
Input 1 for equal, 2 for unequal weighting of D variables ([1:2], <cr>=1):
Reading data file ...
Number of records in data file: 4406
Length of longest entry in data file: 9
Checking for missing values ...
Finished checking
Assigning integer codes to values of 9 categorical variables
Re-checking data ...
Assigning codes to missing values, if any ...
Data checks complete
Normalizing data
Rereading data ...
PCA can be used for variable selection
Do not use PCA if differential item functioning (DIF) scores are wanted
Input 1 to use PCA, 2 otherwise ([1:2], <cr>=2):
#cases w/ miss. D = number of cases with all D values missing

<table>
<thead>
<tr>
<th>Total</th>
<th>#cases w/ #missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>4406</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Number of cases used for training: 4406
Number of split variables: 13
Finished reading data file
Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):
Input file name to store LaTeX code (use .tex as suffix): mult.tex
You can store the variables and/or values used to split and fit in a file
Choose 1 to skip this step, 2 to store split and fit variables,
3 to store split variables and their values
Input your choice ([1:3], <cr>=1):
Input 2 to save node IDs of observations, 1 otherwise ([1:2], <cr>=2):
Input name of file to store terminal node ID of each case: mult.nid
Input 2 to save fitted values at each terminal node; 1 otherwise ([1:2], <cr>=2):
Input name of file to store node fitted values: mult.fit
Input 2 to write R function for predicting new cases, 1 otherwise ([1:2], <cr>=2):
Input file name: mult.r
13.2 Contents of mult.out

Multi-response or longitudinal data without T variables
Pruning by cross-validation
Data description file: nmes.dsc
Training sample file: nmes.txt
Missing value code: NA
Records in data file start on line 1
4 N variables changed to S
Number of D variables: 6
Univariate split variable selection method
Mean-squared errors (MSE) are calculated from normalized D variables
D variables equally weighted
Piecewise constant model
Number of records in data file: 4406
Length of longest entry in data file: 9
Model fitted to subset of observations with complete D values
Neither LDA nor PCA used

Summary information for training sample of size 4406

d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators),
s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
m=missing-value flag variable, p=periodic variable, w=weight

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ofp</td>
<td>0.000</td>
<td>89.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ofnp</td>
<td>0.000</td>
<td>104.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>opp</td>
<td>0.000</td>
<td>141.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>opnp</td>
<td>0.000</td>
<td>155.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>emer</td>
<td>0.000</td>
<td>12.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>hosp</td>
<td>0.000</td>
<td>8.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>health</td>
<td>d</td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>numchron</td>
<td>s</td>
<td>8.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>adddiff</td>
<td>c</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>region</td>
<td>c</td>
<td></td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>age</td>
<td>s</td>
<td>6.600</td>
<td>10.90</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>black</td>
<td>c</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>gender</td>
<td>c</td>
<td></td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
13.2 Contents of `mult.out`  

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Value 1</th>
<th>Value 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>married</td>
<td>c</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>school</td>
<td>s</td>
<td>0.000</td>
<td>18.00</td>
</tr>
<tr>
<td>faminc</td>
<td>s</td>
<td>-1.012</td>
<td>54.84</td>
</tr>
<tr>
<td>employed</td>
<td>c</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>privins</td>
<td>c</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>medicaid</td>
<td>c</td>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>

#cases w/ miss. D = number of cases with all D values missing

<table>
<thead>
<tr>
<th>Total #cases w/ #missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>#cases</td>
</tr>
<tr>
<td>4406</td>
</tr>
</tbody>
</table>

Number of cases used for training: 4406
Number of split variables: 13

Constant fitted to cases with missing values in regressor variables
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE’s for pruned tree: 0.2500

No nodewise interaction tests
Split values for N and S variables based on exhaustive search
Maximum number of split levels: 14
Minimum node sample size: 220

Top-ranked variables and 1-df chi-squared values at root node
1 0.6017E+03 numchron
2 0.3823E+03 health
3 0.2025E+03 adddiff
4 0.9838E+02 privins
5 0.6583E+02 region
6 0.5639E+02 age
7 0.5257E+02 medicaid
8 0.5218E+02 school
9 0.3187E+02 gender
10 0.3126E+02 black
11 0.1892E+02 faminc
12 0.1172E+02 married
13 0.6155E+01 employed

Size and CV Loss and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Tnodes</th>
<th>Mean Loss</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median Loss</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>1.593E+03</td>
<td>1.075E+02</td>
<td>4.586E+02</td>
<td>1.393E+03</td>
<td>9.190E+02</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>1.593E+03</td>
<td>1.075E+02</td>
<td>4.586E+02</td>
<td>1.393E+03</td>
<td>9.190E+02</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>1.593E+03</td>
<td>1.075E+02</td>
<td>4.586E+02</td>
<td>1.393E+03</td>
<td>9.190E+02</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>1.593E+03</td>
<td>1.075E+02</td>
<td>4.586E+02</td>
<td>1.393E+03</td>
<td>9.190E+02</td>
</tr>
</tbody>
</table>

Wei-Yin Loh

GUIDE manual
## 13.2 Contents of mult.out

### 13 MULTIRESPONSE: NMES DATA

<table>
<thead>
<tr>
<th>Cases</th>
<th>Node</th>
<th>Total Cases</th>
<th>Node Split</th>
<th>MSE</th>
<th>variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4406</td>
<td>4406</td>
<td>numchron</td>
<td>1.000E+00</td>
<td>numchron</td>
</tr>
<tr>
<td>2T</td>
<td>2523</td>
<td>2523</td>
<td>numchron</td>
<td>5.688E-01</td>
<td>numchron</td>
</tr>
<tr>
<td>3</td>
<td>1883</td>
<td>1883</td>
<td>health</td>
<td>1.528E+00</td>
<td>-</td>
</tr>
<tr>
<td>6T</td>
<td>426</td>
<td>426</td>
<td>privins</td>
<td>2.282E+00</td>
<td>privins</td>
</tr>
<tr>
<td>7T</td>
<td>1457</td>
<td>1457</td>
<td>privins</td>
<td>1.277E+00</td>
<td>privins</td>
</tr>
</tbody>
</table>

Number of terminal nodes of final tree: 3
Total number of nodes of final tree: 5
Second best split variable (based on curvature test) at root node is health

Regression tree for multi-response data:
For categorical variable splits, values not in training data go to the right

Node 1: numchron <= 1.5000000
  Node 2: Mean cost = 0.56857139
Node 1: numchron > 1.5000000 or NA
  Node 3: health = "poor"
    Node 6: Mean cost = 2.2768607
    Node 3: health /= "poor"
13.2 Contents of mult.out

Node 7: Mean cost = 1.2765754

***************************************************************

Node 1: Intermediate node
A case goes into Node 2 if numchron <= 1.5000000
numchron mean = 1.5419882
Means of ofp, ofnp, opp, opnp, emer, and hosp
5.7744E+00 1.6180E+00 7.5079E-01 5.3609E-01 2.6350E-01
2.9596E-01
----------------------------

Node 2: Terminal node
Means of ofp, ofnp, opp, opnp, emer, and hosp
4.4392E+00 1.4491E+00 4.6968E-01 3.9516E-01 1.6488E-01
1.6647E-01
----------------------------

Node 3: Intermediate node
A case goes into Node 6 if health = "poor"
health mode = "average"
----------------------------

Node 6: Terminal node
Means of ofp, ofnp, opp, opnp, emer, and hosp
9.4319E+00 1.5000E+00 1.5282E+00 6.8310E-01 7.2066E-01
7.9108E-01
----------------------------

Node 7: Terminal node
Means of ofp, ofnp, opp, opnp, emer, and hosp
7.0172E+00 1.9451E+00 1.0103E+00 7.3713E+00 3.0062E-01
3.7543E-01
----------------------------

Case and node IDs are in file: mult.nid
Node fitted values are in file: mult.fit
LaTeX code for tree is in mult.tex
R code is stored in mult.r

The tree is shown in Figure 31. The file mult.fit saves the mean values of the dependent variables in each terminal node:

<table>
<thead>
<tr>
<th>node</th>
<th>ofp</th>
<th>ofnp</th>
<th>opp</th>
<th>opnp</th>
<th>emer</th>
<th>hosp</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.44392E+01</td>
<td>0.14491E+01</td>
<td>0.46968E+00</td>
<td>0.39516E+00</td>
<td>0.16488E+00</td>
<td>0.16647E+00</td>
</tr>
<tr>
<td>6</td>
<td>0.94319E+01</td>
<td>0.15000E+01</td>
<td>0.15282E+01</td>
<td>0.68310E+00</td>
<td>0.72066E+00</td>
<td>0.79108E+00</td>
</tr>
<tr>
<td>7</td>
<td>0.70172E+01</td>
<td>0.19451E+01</td>
<td>0.10103E+01</td>
<td>0.73713E+00</td>
<td>0.30062E+00</td>
<td>0.37543E+00</td>
</tr>
</tbody>
</table>

The file mult.nid gives the terminal node number for each observation, including
13.2 Contents of `mult.out`

Figure 31: GUIDE v.41.1 0.250-SE regression tree for predicting response variables \texttt{ofp}, \texttt{ofnp}, \texttt{opp}, \texttt{opnp}, \texttt{emer}, and \texttt{hosp}, without using PCA at each node. At each split, an observation goes to the left branch if and only if the condition is satisfied. Sample size (in \textit{italics}) and predicted values of \texttt{ofp}, \texttt{ofnp}, \texttt{opp}, \texttt{opnp}, \texttt{emer}, and \texttt{hosp} printed below nodes. Second best split variable at root node is \texttt{health}.
those that are not used to construct the tree (indicated by the letter “n” in the train column of the file).

### 14 Longitudinal response with varying times

The data come from a longitudinal study on the hourly wage of 888 male high-school dropouts (246 black, 204 Hispanic, 438 white), where the observation time points as well as their number (1–13) varied across individuals (Murnane et al., 1999; Singer and Willett, 2003). An earlier version of GUIDE was used to analyze the data in Loh and Zheng (2013).

The response variable is hourly wage (in 1990 dollars) and the predictor variables are hgc (highest grade completed; 6–12), exper (years in labor force; 0.001–12.7 yrs), and race (Black, Hispanic, and White). The data file wagedat.txt is in wide format, where each record refers to one individual. The description file wagedsc.txt is given below. Observation time points are indicated by t. The d and t variable columns may appear anywhere in the data, but the first d must be associated with the first t, second d with the second t, and so on. The number of d and t variables must be the same. Missing d values are permitted to allow for observations with unequal numbers of observation times. Observations with missing values in one or more t variable are excluded from model fitting.

```
wagedat.txt
NA
1
1 id x
2 hgc n
3 exper1 t
4 exper2 t
5 exper3 t
6 exper4 t
7 exper5 t
8 exper6 t
9 exper7 t
10 exper8 t
11 exper9 t
12 exper10 t
13 exper11 t
14 exper12 t
15 exper13 t
16 postexp1 x
17 postexp2 x
```
18 postexp3 x
19 postexp4 x
20 postexp5 x
21 postexp6 x
22 postexp7 x
23 postexp8 x
24 postexp9 x
25 postexp10 x
26 postexp11 x
27 postexp12 x
28 postexp13 x
29 wage1 d
30 wage2 d
31 wage3 d
32 wage4 d
33 wage5 d
34 wage6 d
35 wage7 d
36 wage8 d
37 wage9 d
38 wage10 d
39 wage11 d
40 wage12 d
41 wage13 d
42 ged1 x
43 ged2 x
44 ged3 x
45 ged4 x
46 ged5 x
47 ged6 x
48 ged7 x
49 ged8 x
50 ged9 x
51 ged10 x
52 ged11 x
53 ged12 x
54 ged13 x
55 uerate1 x
56 uerate2 x
57 uerate3 x
58 uerate4 x
59 uerate5 x
60 uerate6 x
61 uerate7 x
62 uerate8 x
63 uerate9 x
14.1 Input file creation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   1
   Name of batch input file: wage.in
   Input 1 for model fitting, 2 for importance or DIF scoring,
   3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: wage.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1): 2
   Choose type of regression model:
   1=linear, 2=quantile, 3=Poisson, 4=censored response,
   5=multiresponse or itemresponse, 6=longitudinal data (with T variables),
   7=binary logistic regression.
   Input choice ([1:7], <cr>=1): 6
   Input 1 for lowess smoothing, 2 for spline smoothing ([1:2], <cr>=1):
   Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
   Input name of data description file (max 100 characters);
   enclose with matching quotes if it has spaces: wagedsc.txt
Reading data description file ...
Training sample file: wagedat.txt
Missing value code: NA
Records in data file start on line 1
One N variable changed to S
Number of D variables: 13
D variables are:
wage1
wage2
wage3
wage4
wage5
wage6
wage7
wage8
wage9
14.1 Input file creation

T variables are:
- exper1
- exper2
- exper3
- exper4
- exper5
- exper6
- exper7
- exper8
- exper9
- exper10
- exper11
- exper12
- exper13

D variables can be grouped into segments to look for patterns
Input 1 for equal-sized groups, 2 for custom groups ([1:2], <cr>=1):
Input number of roughly equal-sized groups ([2:9], <cr>=3):
Input number of interpolating points for prediction ([10:100], <cr>=31):

Reading data file ...
Number of records in data file: 888
Length of longest entry in data file: 16
Checking for missing values ...
Finished checking
Missing values found in D variables
Assigning integer codes to values of 1 categorical variables
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...
Data checks complete
Creating missing value indicators ...
Rereading data ...

#cases w/ miss. D = number of cases with all D values missing
Total #cases w/ #missing
#cases miss. D ord. vals #X-var #N-var #F-var #S-var
888 0 0 40 0 0 1
#P-var #M-var #B-var #C-var #I-var
0 0 0 1 0

Number of cases used for training: 888
Number of split variables: 2
Number of cases excluded due to 0 W or missing D variable: 0
Finished reading data file
Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):
14.2 Contents of wage.out

Longitudinal data with T variables
Lowess smoothing
Pruning by cross-validation
Data description file: wagedsc.txt
Training sample file: wagedat.txt
Missing value code: NA
Records in data file start on line 1
One N variable changed to S
Number of D variables: 13
Number of D variables: 13
D variables are:
wage1
wage2
wage3
wage4
wage5
wage6
wage7
wage8
wage9
wage10
wage11
wage12
wage13
T variables are:
exper1
14.2 Contents of wage.out

Number of records in data file: 888
Length of longest entry in data file: 16
Missing values found in D variables
Model fitted to subset of observations with complete D values

Summary information for training sample of size 888

d=dependent, b=split and fit cat variable using indicator variables, 
c=split-only categorical, i=fit-only categorical (via indicators), 
s=split-only numerical, n=split and fit numerical, f=fit-only numerical, 
m=missing-value flag variable, p=periodic variable, w=weight

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>hgc</td>
<td>6.000</td>
<td>12.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>exper1</td>
<td>0.000</td>
<td>5.637</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>exper2</td>
<td>0.000</td>
<td>7.584</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>exper3</td>
<td>0.000</td>
<td>9.777</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>exper4</td>
<td>0.000</td>
<td>10.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>exper5</td>
<td>0.000</td>
<td>11.78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>exper6</td>
<td>0.000</td>
<td>10.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>exper7</td>
<td>0.000</td>
<td>11.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>exper8</td>
<td>0.000</td>
<td>10.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>exper9</td>
<td>0.000</td>
<td>11.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>exper10</td>
<td>0.000</td>
<td>12.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>exper11</td>
<td>0.000</td>
<td>11.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>exper12</td>
<td>0.000</td>
<td>12.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>exper13</td>
<td>0.000</td>
<td>12.70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>wage1</td>
<td>2.030</td>
<td>68.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>wage2</td>
<td>2.069</td>
<td>50.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>wage3</td>
<td>2.046</td>
<td>34.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>wage4</td>
<td>2.117</td>
<td>33.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>wage5</td>
<td>2.104</td>
<td>49.30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>wage6</td>
<td>2.208</td>
<td>74.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>wage7</td>
<td>2.104</td>
<td>47.28</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
14.2 Contents of wage.out

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>36</td>
<td>wage8</td>
<td>d</td>
<td>2.316</td>
<td>37.71</td>
</tr>
<tr>
<td>37</td>
<td>wage9</td>
<td>d</td>
<td>2.529</td>
<td>46.11</td>
</tr>
<tr>
<td>38</td>
<td>wage10</td>
<td>d</td>
<td>2.998</td>
<td>56.54</td>
</tr>
<tr>
<td>39</td>
<td>wage11</td>
<td>d</td>
<td>4.084</td>
<td>22.20</td>
</tr>
<tr>
<td>40</td>
<td>wage12</td>
<td>d</td>
<td>3.432</td>
<td>46.20</td>
</tr>
<tr>
<td>41</td>
<td>wage13</td>
<td>d</td>
<td>4.563</td>
<td>7.776</td>
</tr>
<tr>
<td>68</td>
<td>race</td>
<td>c</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

Total # cases w/ # missing:
<table>
<thead>
<tr>
<th># cases</th>
<th>miss. D</th>
<th>ord. vals</th>
<th>X-var</th>
<th>N-var</th>
<th>F-var</th>
<th>S-var</th>
<th>P-var</th>
<th>M-var</th>
<th>B-var</th>
<th>C-var</th>
<th>I-var</th>
</tr>
</thead>
<tbody>
<tr>
<td>888</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Number of cases used for training: 888
Number of split variables: 2
Number of cases excluded due to 0 W or missing D variable: 0

Constant fitted to cases with missing values in regressor variables
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE's for pruned tree: 0.2500

No nodewise interaction tests
Split values for N and S variables based on exhaustive search
Maximum number of split levels: 10
Minimum node sample size: 44
Top-ranked variables and 1-df chi-squared values at root node
1 0.1235E+02 hgc
2 0.6915E+01 race

Size and CV Loss and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Tnodes</th>
<th>Mean Loss</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median Loss</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>1.262E+02</td>
<td>1.042E+01</td>
<td>9.660E+00</td>
<td>1.244E+02</td>
<td>1.005E+01</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>1.262E+02</td>
<td>1.042E+01</td>
<td>9.660E+00</td>
<td>1.244E+02</td>
<td>1.005E+01</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>1.243E+02</td>
<td>1.054E+01</td>
<td>9.934E+00</td>
<td>1.206E+02</td>
<td>1.029E+01</td>
</tr>
<tr>
<td>4*</td>
<td>3</td>
<td>1.235E+02</td>
<td>1.051E+01</td>
<td>9.863E+00</td>
<td>1.205E+02</td>
<td>1.077E+01</td>
</tr>
<tr>
<td>5+</td>
<td>2</td>
<td>1.237E+02</td>
<td>1.060E+01</td>
<td>1.006E+01</td>
<td>1.204E+02</td>
<td>1.102E+01</td>
</tr>
<tr>
<td>6**</td>
<td>1</td>
<td>1.244E+02</td>
<td>1.065E+01</td>
<td>1.011E+01</td>
<td>1.210E+02</td>
<td>1.171E+01</td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 3 terminal nodes
0-SE tree based on median is marked with + and has 2 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
** tree same as ++ tree
** tree same as -- tree

Wei-Yin Loh 261 GUIDE manual
++ tree same as -- tree

WARNING: tree based on mean CV estimate of error has no splits
Choosing smallest nontrivial tree with no larger CV error estimate

Structure of final tree. Each terminal node is marked with a T.

Cases fit give the number of cases used to fit node
MSE is residual sum of squares divided by number of cases in node

<table>
<thead>
<tr>
<th>Node</th>
<th>Total Cases</th>
<th>Cases fit</th>
<th>MSE</th>
<th>Split</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>888</td>
<td>888</td>
<td>1.222E+02</td>
<td>hgc</td>
</tr>
<tr>
<td>2T</td>
<td>577</td>
<td>577</td>
<td>1.040E+02</td>
<td>race</td>
</tr>
<tr>
<td>3T</td>
<td>311</td>
<td>311</td>
<td>1.513E+02</td>
<td>race</td>
</tr>
</tbody>
</table>

Number of terminal nodes of final tree: 2
Total number of nodes of final tree: 3
Second best split variable (based on curvature test) at root node is race

Regression tree for longitudinal data:

Node 1: hgc <= 9.5000000
Node 2: Mean cost = 103.80991
Node 1: hgc > 9.5000000 or NA
Node 3: Mean cost = 150.79730

***************************************************************

Node 1: Intermediate node
A case goes into Node 2 if hgc <= 9.5000000
hgc mean = 8.9166667

Node 2: Terminal node

Node 3: Terminal node

Case and node IDs are in file: wage.nid
Node fitted values are in file: wage.fit
LaTeX code for tree is in wage.tex
R code is stored in wage.r
Split and fit variable names are stored in wage.var

Figure 32 shows the tree and Figure 33 plots lowess-smoothed curves of mean wage in the two terminal nodes. The figure is produced by the following R code.

z <- read.table("wagedat.txt",header=FALSE)
Figure 32: GUIDE v.41.1 0.053-SE (0.250-SE has no splits) regression tree for predicting longitudinal variables \textit{wage1}, \textit{wage2}, etc. At each split, an observation goes to the left branch if and only if the condition is satisfied. Sample size \textit{(in italics)} printed below nodes. Second best split variable at root node is \textit{race}.

Figure 33: Lowess-smoothed mean wage curves in the terminal nodes of Figure 32.
14.2 Contents of wage.out

`names(z) <- c("id","hgc","exper1","exper2","exper3","exper4","exper5","exper6",
    "exper7","exper8","exper9","exper10","exper11","exper12","exper13",
    "postexp1","postexp2","postexp3","postexp4","postexp5","postexp6",
    "postexp7","postexp8","postexp9","postexp10","postexp11","postexp12",
    "postexp13","wage1","wage2","wage3","wage4","wage5","wage6","wage7",
    "wage8","wage9","wage10","wage11","wage12","wage13","ged1","ged2",
    "ged3","ged4","ged5","ged6","ged7","ged8","ged9","ged10","ged11",
    "ged12","ged13","uerate1","uerate2","uerate3","uerate4","uerate5",
    "uerate6","uerate7","uerate8","uerate9","uerate10","uerate11",
    "uerate12","uerate13","race")`

`exper <- c(z$exper1,z$exper2,z$exper3,z$exper4,z$exper5,z$exper6,z$exper7,
    z$exper8,z$exper9,z$exper10,z$exper11,z$exper12,z$exper13)`

`wage <- c(z$wage1,z$wage2,z$wage3,z$wage4,z$wage5,z$wage6,z$wage7,z$wage8,
    z$wage9,z$wage10,z$wage11,z$wage12,z$wage13)`

`xr <- range(exper,na.rm=TRUE)`

`yr <- range(wage,na.rm=TRUE)`

`guide.fit <- read.table("wage.fit",header=TRUE)`

`g.node <- guide.fit$node`  
`g.start <- guide.fit$t.start`  
`g.end <- guide.fit$t.end`  
`n <- length(g.node)`  
`m <- dim(guide.fit)[2]`  
`npts <- m-3 # number of time points for plotting`

`xvals <- guide.fit[,2:3]`  
`xvals <- as.numeric(unlist(xvals))`  
`yvals <- guide.fit[,4:m]`  
`yvals <- as.numeric(unlist(yvals))`  
`plot(range(xvals),range(yvals),type="n",xlab="exper (years)",ylab="hourly wage ($)")`  
`leg.col <- c("blue","red")`  
`leg.lty <- c(1,2)`  
`for(i in 1:n){`  
`    node <- g.node[i]`  
`    start <- g.start[i]`  
`    end <- g.end[i]`  
`    gap <- (end-start)/(npts-1)`  
`    x <- start+(0:(npts-1))*gap`  
`    y <- as.numeric(guide.fit[i,4:m])`  
`    lines(x,y,col=leg.col[i],lty=leg.lty[i])`  
`}`  
`leg.txt <- c(expression(paste("hgc" <= 9)),expression(paste("hgc" > 9)))`  
`legend("topleft",legend=leg.txt,lty=leg.lty,col=leg.col,lwd=2)`

The plotting values are obtained from the result file wage.fit whose contents are
given below. The first column gives the node number and the next two columns the start and end of the times at which fitted values are computed. The other columns give the fitted values equally spaced between the start and end times.

<table>
<thead>
<tr>
<th>node</th>
<th>t.start</th>
<th>t.end</th>
<th>fitted1</th>
<th>fitted2</th>
<th>fitted3</th>
<th>fitted4</th>
<th>fitted5</th>
<th>fitted6</th>
<th>fitted7</th>
<th>fitted8</th>
<th>fitted9</th>
<th>fitted10</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.10000E-02</td>
<td>0.12700E+02</td>
<td>0.48875E+01</td>
<td>0.51221E+01</td>
<td>0.53241E+01</td>
<td>0.54668E+01</td>
<td>0.55738E+01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.20000E-02</td>
<td>0.12558E+02</td>
<td>0.57699E+01</td>
<td>0.58884E+01</td>
<td>0.60035E+01</td>
<td>0.60997E+01</td>
<td>0.63206E+01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The contents of the file wage.var are given below. The 1st column gives the node number. The 2nd column is a letter, with t indicating that the node is terminal and c, s, or n indicating an intermediate node split on a c, n or s variable. The 3rd column gives the name of the variable used to split the node; the name NONE is used if a terminal node cannot be split by any variable. The 4th column gives the name of the interacting variable if there is one; otherwise the name of the split variable is repeated. If the node is terminal, the 5th column contains the letter “t”; otherwise if it is non-terminal, the 5th column is an integer indicating the number of split values to follow (a split on a c variable may have more than one value). In the example below, node 1 is split on s variable hgc at value 9.50. Nodes 2 and 3 are terminal nodes; each would be split on race if they were not terminal.

1 s hgc hgc 1 0.9500000000E+01
2 t race race t
3 t race race t

15 Logistic regression

If the dependent variable $Y$ takes values 0 and 1, GUIDE can construct a tree model such that a simple or multiple linear logistic regression model is fitted in each node. The tree model may be more efficient (in terms of size and prediction accuracy) if a preliminary estimate of $p = P(Y = 1)$ is available. The preliminary estimate of $p$ is not necessary, but it may be easily obtained by fitting a GUIDE forest or kernel discriminant model to the data. If a variable containing the estimated $p$ values is included in the data, it should be specified as an “e” variable in the description file (see Section 3.1). Missing values in the predictor variables used in the logistic regression node models are imputed with node means; see Loh (2021) for more details.

We use the NHTSA data to demonstrate this, with $Y = \text{HIC2}$, which takes value 1 if HIC > 999 and 0 otherwise. The description file is nhtsadsc2.txt. The “e” variable is estHIC2 which is a column of estimated values of $p = P(Y = 1)$ obtained from GUIDE forest.
15.1 Piecewise constant

15.1.1 Input file creation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: logitc.in
   Input 1 for model fitting, 2 for importance or DIF scoring,
   3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: logitc.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1): 2
   Choose type of regression model:
     1=linear, 2=quantile, 3=Poisson, 4=censored response,
     5=multiprespon or itemresponse, 6=longitudinal data (with T variables),
     7=binary logistic regression.
   Input choice ([1:7], <cr>=1): 7
   Choose complexity of model to use at each node:
   Choose 1 for multiple regression (recommended for prediction)
   Choose 2 for best simple polynomial in one N or F variable
   Choose 3 for constant fit (recommended for interpretability or if there is an R variable)
   Input 1 for default options, 2 otherwise ([1:2], <cr>=3):
   Input name of data description file (max 100 characters);
   enclose with matching quotes if it has spaces: nhtsadsc2.txt
   Reading data description file ...
   Training sample file: nhtsadatam.txt
   Missing value code: NA
   Records in data file start on line 2
   48 N variables changed to S
   Warning: B variables changed to C
   D variable is HIC2
   Reading data file ...
   Number of records in data file: 3310
   Length of longest entry in data file: 19
   Checking for missing values ...
   Finished checking
   Missing values found in D variable
   Missing values found among categorical variables
   Separate categories will be created for missing categorical variables
   Missing values found among non-categorical variables
   Assigning integer codes to values of 13 categorical variables
   Finished assigning codes to 10 categorical variables
   Associating missing values of N, P and S variables with M variable codes ...

Wei-Yin Loh 266 GUIDE manual
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...
Data checks complete
Creating missing value indicators ...
Rereading data ...

<table>
<thead>
<tr>
<th>#cases</th>
<th>miss. D</th>
<th>ord. vals</th>
<th>#X-var</th>
<th>#N-var</th>
<th>#F-var</th>
<th>#S-var</th>
</tr>
</thead>
<tbody>
<tr>
<td>3310</td>
<td>34</td>
<td>3310</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>48</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#P-var</th>
<th>#M-var</th>
<th>#B-var</th>
<th>#C-var</th>
<th>#I-var</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>42</td>
<td>0</td>
<td>13</td>
<td>0</td>
</tr>
</tbody>
</table>

Number of cases used for training: 3276
Number of split variables: 61
Number of cases excluded due to 0 W or missing D variable: 34
Proportion of ones in HIC2 variable: 8.4554334554334559E-002

Finished reading data file
Minimum number of D=0 and D=1 in each node: 9
Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2): 
Input file name to store LaTeX code (use .tex as suffix): logitc.tex
You can store the variables and/or values used to split and fit in a file
Choose 1 to skip this step, 2 to store split and fit variables, 3 to store split variables and their values
Input your choice ([1:3], <cr>=1):
Input 2 to save fitted values and node IDs, 1 otherwise ([1:2], <cr>=2):
Input name of file to store node ID and fitted value of each case: logitc.fit
Input 2 to write R function for predicting new cases, 1 otherwise ([1:2], <cr>=2):
Input file name: logitc.r
Input rank of top variable to split root node ([1:67], <cr>=1):
Input file is created!

15.1.2 Contents of logitc.out

Binary logistic regression tree
Pruning by cross-validation
Data description file: nhtsadsc2.txt
Training sample file: nhtsadatam.txt
Missing value code: NA
Records in data file start on line 2
48 N variables changed to S
Warning: B variables changed to C
D variable is HIC2
Piecewise constant model
Number of records in data file: 3310
Length of longest entry in data file: 19
Missing values found in D variable
15.1 Piecewise constant

Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables

Summary information for training sample of size 3276 (excluding observations with non-positive weight or missing values in $d$, $e$, $t$, $r$ or $z$ variables)

$d=$dependent, $b=$split and fit cat variable using indicator variables,
$c=$split-only categorical, $i=$fit-only categorical (via indicators),
$s=$split-only numerical, $n=$split and fit numerical, $f=$fit-only numerical,
$m=$missing-value flag variable, $p=$periodic variable, $w=$weight,
$e=$estimated success probability

Levels of $M$ variables are for missing values in associated variables

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BARR1G</td>
<td>c</td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>BARSHP</td>
<td>c</td>
<td></td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>BARANG</td>
<td>p</td>
<td>0.000</td>
<td>330.0</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>BARDIA</td>
<td>s</td>
<td>191.0</td>
<td>1000.</td>
<td>2807</td>
</tr>
<tr>
<td>5</td>
<td>OCCWT</td>
<td>s</td>
<td>72.00</td>
<td>83.00</td>
<td>3265</td>
</tr>
<tr>
<td>6</td>
<td>OCCWT_</td>
<td>m</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>104</td>
<td>VEHSPD</td>
<td>s</td>
<td>0.3000</td>
<td>99.10</td>
<td>6</td>
</tr>
<tr>
<td>105</td>
<td>VEHSPD_</td>
<td>m</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>106</td>
<td>CRBANG</td>
<td>p</td>
<td>0.000</td>
<td>315.0</td>
<td>24</td>
</tr>
<tr>
<td>107</td>
<td>PDOF</td>
<td>p</td>
<td>0.000</td>
<td>345.0</td>
<td>23</td>
</tr>
<tr>
<td>108</td>
<td>CARANG</td>
<td>p</td>
<td>0.000</td>
<td>99.00</td>
<td>991</td>
</tr>
<tr>
<td>109</td>
<td>VEHOR</td>
<td>p</td>
<td>0.000</td>
<td>90.00</td>
<td>995</td>
</tr>
<tr>
<td>110</td>
<td>RSTFRT</td>
<td>c</td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>111</td>
<td>HIC2</td>
<td>d</td>
<td>0.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>112</td>
<td>estHIC2</td>
<td>e</td>
<td>0.000</td>
<td>0.8455</td>
<td></td>
</tr>
</tbody>
</table>

Total #cases w/ #missing
#cases miss. D ord. vals #X-var #N-var #F-var #S-var
3310 34 3310 1 0 0 48
#P-var #M-var #B-var #C-var #I-var
6 42 0 13 0

Number of cases used for training: 3276
Number of split variables: 61
Number of cases excluded due to 0 W or missing D variable: 34
Proportion of ones in HIC2 variable: 0.084554

Constant fitted to cases with missing values in regressor variables
Pruning by $v$-fold cross-validation, with $v = 10$
Selected tree is based on mean of CV estimates
Number of SE's for pruned tree: 0.2500
15.1 Piecewise constant

Nodewise interaction tests on all variables
Fraction of cases used for splitting each node: 1.0000
Maximum number of split levels: 13
Minimum node sample size: 65
Minimum number of D=0 and D=1 in each node: 9
Top-ranked variables and 1-df chi-squared values at root node

1 0.1218E+04 COLMEC
2 0.9001E+03 YEAR
3 0.8714E+03 MODELD
4 0.7917E+03 RSTFRT
5 0.6935E+03 HS
6 0.5377E+03 HR
7 0.3959E+03 CS

Size and CV Loss and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Nodes</th>
<th>Mean Loss</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median Loss</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-</td>
<td>8</td>
<td>4.534E-01</td>
<td>2.045E-02</td>
<td>7.571E-03</td>
<td>4.498E-01</td>
<td>7.411E-03</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>4.599E-01</td>
<td>2.042E-02</td>
<td>6.132E-03</td>
<td>4.516E-01</td>
<td>7.028E-03</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>4.580E-01</td>
<td>2.012E-02</td>
<td>6.699E-03</td>
<td>4.516E-01</td>
<td>7.863E-03</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4.580E-01</td>
<td>2.012E-02</td>
<td>6.699E-03</td>
<td>4.516E-01</td>
<td>7.863E-03</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>4.580E-01</td>
<td>2.012E-02</td>
<td>6.699E-03</td>
<td>4.516E-01</td>
<td>7.863E-03</td>
</tr>
<tr>
<td>6**</td>
<td>2</td>
<td>4.580E-01</td>
<td>2.012E-02</td>
<td>6.699E-03</td>
<td>4.516E-01</td>
<td>7.863E-03</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>5.795E-01</td>
<td>2.316E-02</td>
<td>2.216E-03</td>
<td>5.834E-01</td>
<td>3.465E-03</td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 8 terminal nodes
0-SE tree based on median is marked with + and has 8 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
* tree same as + tree
** tree same as ++ tree
* tree same as -- tree

Following tree is based on mean CV with naive SE estimate (**)

Structure of final tree. Each terminal node is marked with a T.

D-mean is mean of HIC2 in the node
Cases fit give the number of cases used to fit node
Node deviance is residual deviance divided by residual degrees of freedom

<table>
<thead>
<tr>
<th>Node</th>
<th>Total Cases</th>
<th>Matrix Node</th>
<th>Node Split</th>
<th>Other variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>label</td>
<td>cases fit</td>
<td>rank D-mean</td>
<td>deviance variable</td>
<td>variables</td>
</tr>
</tbody>
</table>

Wei-Yin Loh 269 GUIDE manual
Regression tree:
For categorical variable splits, values not in training data go to the right

Node 1: COLMEC = "BWU", "CYL", "NA", "NAP", "UNK"
   Node 2: HIC2 proportion of 1s = 0.27969349E-1
   Node 1: COLMEC /= "BWU", "CYL", "NA", "NAP", "UNK"
   Node 3: HIC2 proportion of 1s = 0.30630631

***************************************************************
Predictor means below are means of cases with no missing values.

WARNING: p-values below not adjusted for split search. For a bootstrap solution see:

Node 1: Intermediate node
A case goes into Node 2 if COLMEC = "BWU", "CYL", "NA", "NAP", "UNK"
   COLMEC mode = "UNK"
   Coefficients of logit function
   Regressor    Coefficient    t-stat    p-value
   Constant   -2.382        17.39    0.000
   Proportion of ones in variable HIC2 = 0.845543E-1

Node 2: Terminal node
   Coefficients of logit:
   Regressor    Coefficient    t-stat    p-value
   Constant   -3.548        8.666    0.000
   Proportion of ones in variable HIC2 = 0.279693E-1

Node 3: Terminal node
   Coefficients of logit:
Figure 34: GUIDE v.41.1 0.250-SE piecewise-constant logistic regression tree for predicting $P(\text{HIC2}=1)$. At each split, an observation goes to the left branch if and only if the condition is satisfied. $S_1 = \{\text{BWU, CYL, NA, NAP, UNK}\}$. Sample size (in italics) and proportion of 1s in HIC2 printed below nodes. Terminal nodes with proportions of 1s above and below value of 0.08 at root node are colored yellow and vermilion respectively. Second best split variable at root node is \textsc{year}.

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.8174</td>
<td>17.15</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Proportion of ones in variable HIC2 = 0.306306

---

Observed and fitted values are stored in logitc.fit
Latex code for tree is in logitc.tex
R code is stored in logitc.r

The logistic regression tree is shown in Figure 34.

15.2 Simple linear

We can also construct a logistic regression tree with a simple linear logistic regression model fitted to each node.

15.2.1 Input file creation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: logits.in
   Input 1 for model fitting, 2 for importance or DIF scoring,
   3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: logits.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):

Wei-Yin Loh 271 GUIDE manual
Input 1 for classification, 2 for regression, 3 for propensity score grouping
Input your choice ([1:3], <cr>=1): 2
Choose type of regression model:
  1=linear, 2=quantile, 3=Poisson, 4=censored response,
  5=multiresponse or itemresponse, 6=longitudinal data (with T variables),
  7=binary logistic regression.
Input choice ([1:7], <cr>=1): 7
Choose complexity of model to use at each node:
Choose 1 for multiple regression (recommended for prediction)
Choose 2 for best simple polynomial in one N or F variable
Choose 3 for constant fit (recommended for interpretability or if there is an R variable)
1: multiple linear, 2: best simple polynomial, 3: constant ([1:3], <cr>=3): 2
Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
Input name of data description file (max 100 characters);
enclose with matching quotes if it has spaces: nhtsadsc2.txt
Reading data description file ...
Training sample file: nhtsadatam.txt
Missing value code: NA
Records in data file start on line 2
Warning: B variables changed to C
D variable is HIC2
Reading data file ...
Number of records in data file: 3310
Length of longest entry in data file: 19
Checking for missing values ...
Finished checking
Missing values found in D variable
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Assigning integer codes to values of 13 categorical variables
Finished assigning codes to 10 categorical variables
Associating missing values of N, P and S variables with M variable codes ...
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...
Data checks complete
Creating missing value indicators ...
Rereading data ...

<table>
<thead>
<tr>
<th>Total</th>
<th>#cases w/</th>
<th>#missing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cases</td>
<td>miss. D</td>
</tr>
<tr>
<td>3310</td>
<td>34</td>
<td>3310</td>
</tr>
<tr>
<td>6</td>
<td>42</td>
<td>0</td>
</tr>
</tbody>
</table>

Number of cases used for training: 3276
Number of split variables: 61

Wei-Yin Loh

GUIDE manual
15.2 Simple linear

15.2.2 Contents of logits.out

Binary logistic regression tree
Pruning by cross-validation
Data description file: nhtsadsc2.txt
Training sample file: nhtsadatam.txt
Missing value code: NA
Records in data file start on line 2
Warning: B variables changed to C
D variable is HIC2
Piecewise simple linear logistic model
Number of records in data file: 3310
Length of longest entry in data file: 19
Missing values found in D variable
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables

Summary information for training sample of size 3276 (excluding observations with non-positive weight or missing values in d, e, t, r or z variables)
d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators),
as=split-only numerical, n=split and fit numerical, f=fit-only numerical,
m=missing-value flag variable, p=periodic variable, w=weight,
e=estimated success probability
Levels of M variables are for missing values in associated variables
15.2 Simple linear logistic regression

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BARRIG</td>
<td>c</td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>BARSHP</td>
<td>c</td>
<td></td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>BARANG</td>
<td>p</td>
<td>0.000</td>
<td>330.0</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>BARDIA</td>
<td>n</td>
<td>1.9100E+02</td>
<td>1000.0</td>
<td>2807</td>
</tr>
<tr>
<td>5</td>
<td>OCCWT</td>
<td>n</td>
<td>7.2000E+01</td>
<td>83.00</td>
<td>3265</td>
</tr>
<tr>
<td>6</td>
<td>OCCWT_</td>
<td>m</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>DUMSIZ</td>
<td>c</td>
<td></td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>HH</td>
<td>n</td>
<td>5.8000E+01</td>
<td>4321.0</td>
<td>150</td>
</tr>
<tr>
<td>9</td>
<td>HH_</td>
<td>m</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>VEHSPD</td>
<td>n</td>
<td>3.0000E-01</td>
<td>99.10</td>
<td>6</td>
</tr>
<tr>
<td>11</td>
<td>VEHSPD_</td>
<td>m</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>CRBANG</td>
<td>p</td>
<td>0.000</td>
<td>315.0</td>
<td>24</td>
</tr>
<tr>
<td>13</td>
<td>PDOF</td>
<td>p</td>
<td>0.000</td>
<td>345.0</td>
<td>23</td>
</tr>
<tr>
<td>14</td>
<td>CARANG</td>
<td>p</td>
<td>0.000</td>
<td>99.00</td>
<td>991</td>
</tr>
<tr>
<td>15</td>
<td>VEHOR</td>
<td>p</td>
<td>0.000</td>
<td>90.00</td>
<td>995</td>
</tr>
<tr>
<td>16</td>
<td>RSTFRT</td>
<td>c</td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>HIC2</td>
<td>d</td>
<td>0.000</td>
<td>1.000</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>estHIC2</td>
<td>e</td>
<td>0.000</td>
<td>0.8455</td>
<td></td>
</tr>
</tbody>
</table>

Total #cases w/ #missing

<table>
<thead>
<tr>
<th>#cases</th>
<th>miss. D</th>
<th>ord. vals</th>
<th>#X-var</th>
<th>#N-var</th>
<th>#F-var</th>
<th>#S-var</th>
<th>#P-var</th>
<th>#M-var</th>
<th>#B-var</th>
<th>#C-var</th>
<th>#I-var</th>
</tr>
</thead>
<tbody>
<tr>
<td>3310</td>
<td>34</td>
<td>3310</td>
<td>1</td>
<td>48</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>42</td>
<td>0</td>
<td>13</td>
<td>0</td>
</tr>
</tbody>
</table>

Number of cases used for training: 3276
Number of split variables: 61
Number of cases excluded due to 0 W or missing D variable: 34
Proportion of ones in HIC2 variable: 0.084554

Constant fitted to cases with missing values in regressor variables
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE's for pruned tree: 0.2500

Nodewise interaction tests on all variables
Fraction of cases used for splitting each node: 1.0000
Maximum number of split levels: 13
Minimum node sample size: 65
Minimum number of D=0 and D=1 in each node: 9
Top-ranked variables and 1-df chi-squared values at root node
1 0.4911E+03  RSTFRT
2 0.4567E+03  MODELD

Wei-Yin Loh
15.2 Simple linear LOGISTIC REGRESSION

3 0.3172E+03 IMPANG
4 0.2900E+03 COLMEC
5 0.2769E+03 BARDIA
6 0.2617E+03 BARSHP
:
65 0.8221E+00 CARANG
66 0.5257E+00 WHLBAS

Size and CV Loss and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Nodes</th>
<th>Mean Loss</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median Loss</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>6.281E-01</td>
<td>2.156E-02</td>
<td>6.402E-02</td>
<td>5.496E-01</td>
<td>7.028E-02</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>6.509E-01</td>
<td>2.157E-02</td>
<td>6.659E-02</td>
<td>6.092E-01</td>
<td>9.245E-02</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>6.836E-01</td>
<td>2.124E-02</td>
<td>6.729E-02</td>
<td>6.446E-01</td>
<td>1.075E-01</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>6.786E-01</td>
<td>2.106E-02</td>
<td>6.970E-02</td>
<td>5.956E-01</td>
<td>1.181E-01</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>6.774E-01</td>
<td>2.091E-02</td>
<td>6.991E-02</td>
<td>5.896E-01</td>
<td>1.185E-01</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>6.765E-01</td>
<td>2.087E-02</td>
<td>7.003E-02</td>
<td>5.896E-01</td>
<td>1.194E-01</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>7.436E-01</td>
<td>1.937E-02</td>
<td>8.730E-02</td>
<td>6.943E-01</td>
<td>1.526E-01</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>4.547E-01</td>
<td>1.932E-02</td>
<td>9.157E-03</td>
<td>4.653E-01</td>
<td>1.100E-02</td>
</tr>
<tr>
<td>9**</td>
<td>1</td>
<td>4.547E-01</td>
<td>1.932E-02</td>
<td>9.157E-03</td>
<td>4.653E-01</td>
<td>1.100E-02</td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 1 terminal node
0-SE tree based on median is marked with + and has 1 terminal node
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
* tree, ** tree, + tree, and ++ tree all the same

Following tree is based on mean CV with naive SE estimate (**)

Structure of final tree. Each terminal node is marked with a T.

D-mean is mean of HIC2 in the node
Cases fit give the number of cases used to fit node
Node deviance is residual deviance divided by residual degrees of freedom

<table>
<thead>
<tr>
<th>Node</th>
<th>Total Cases</th>
<th>Matrix Node</th>
<th>Split Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>1T</td>
<td>3276</td>
<td>3276</td>
<td>2 8.455E-02 4.546E-01 RSTFRT -YEAR</td>
</tr>
</tbody>
</table>

Best split at root node is on RSTFRT

Number of terminal nodes of final tree: 1
Total number of nodes of final tree: 1
Best split variable (based on curvature test) at root node is RSTFRT

Regression tree:

Node 1: HIC2 proportion of 1s = 0.84554335E-1
Predictor means below are means of cases with no missing values. Regression coefficients are computed from the complete cases.

Node 1: Terminal node
Coefficients of logit:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>258.0</td>
<td>17.26</td>
<td>0.6661E-15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YEAR</td>
<td>-0.1306</td>
<td>-17.38</td>
<td>0.000</td>
<td>1972.</td>
<td>2000.</td>
<td>2017.</td>
</tr>
</tbody>
</table>

If regressors have missing values, predicted value = 0.84554335E-1

Observed and fitted values are stored in logits.fit
Regressor names and coefficients are stored in logits.reg
LaTeX code for tree is in logits.tex
R code is stored in logits.r

The results show that the tree has no splits. It fits a simple linear logistic regression model to the whole data set with \( \text{YEAR} \) as linear predictor. If the value of \( \text{YEAR} \) is missing, the predicted value of \( p \) is the mean of \( \text{HIC2} \).

15.3 Multiple linear

This section shows how to fit a multiple logistic regression to each node of the tree.

15.3.1 Input file creation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: logitm.in
   Input 1 for model fitting, 2 for importance or DIF scoring,
   3 for data conversion ([1:3], <cr>=1):
   Name of batch output file: logitm.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1):
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1): 2
   Choose type of regression model:
   1=linear, 2=quantile, 3=Poisson, 4=censored response,
   5=multiresponse or itemresponse, 6=longitudinal data (with T variables),
   7=binary logistic regression.
   Input choice ([1:7], <cr>=1): 7
   Choose complexity of model to use at each node:
Choose 1 for multiple regression (recommended for prediction)
Choose 2 for best simple polynomial in one N or F variable
Choose 3 for constant fit (recommended for interpretability or if there is an R variable)

1: multiple linear, 2: best simple polynomial in one N or F variable
Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
Input name of data description file (max 100 characters);
enclose with matching quotes if it has spaces: nhtsadsc2.txt

Reading data description file ...
Training sample file: nhtsadatam.txt
Missing value code: NA
Records in data file start on line 2
D variable is HIC2
Reading data file ...
Number of records in data file: 3310
Length of longest entry in data file: 19
Checking for missing values ...
Finished checking
Missing values found in D variable
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Assigning integer codes to values of 13 categorical variables
Finished assigning codes to 10 categorical variables
Associating missing values of N, P and S variables with M variable codes ...
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...
Data checks complete
GUIDE will try to create the variables in the description file.
If it is unsuccessful, please create the columns yourself...
Number of dummy variables created: 20
Creating dummy variables ...
Creating missing value indicators ...
Rereading data ...

Total #cases w/ #missing
#cases miss. D ord. vals #X-var #N-var #F-var #S-var
3310 34 3310 1 48 48 0
#P-var #M-var #B-var #C-var #I-var
6 42 1 12 0
Number of cases used for training: 3276
Number of split variables: 61
Number of dummy variables created: 20
Number of cases excluded due to 0 W or missing D variable: 34
Proportion of ones in HIC2 variable: 8.4554334554334559E-002
Finished reading data file
Minimum number of D=0 and D=1 in each node: 9

Wei-Yin Loh 277

GUIDE manual
15.3.2 Results

Binary logistic regression tree
Pruning by cross-validation
Data description file: nhtsadsc2.txt
Training sample file: nhtsadatam.txt
Missing value code: NA
Records in data file start on line 2
D variable is HIC2
Piecewise multiple linear logistic model
Number of records in data file: 3310
Length of longest entry in data file: 19
Missing values found in D variable
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Number of dummy variables created: 20

Summary information for training sample of size 3276 (excluding observations with non-positive weight or missing values in d, e, t, r or z variables)
d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators),
s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
m=missing-value flag variable, p=periodic variable, w=weight,
e=estimated success probability
Levels of M variables are for missing values in associated variables

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BARRIG</td>
<td>c</td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>BARSHP</td>
<td>b</td>
<td></td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>BARANG</td>
<td>p 0.000</td>
<td>330.0</td>
<td>360</td>
<td>14</td>
</tr>
</tbody>
</table>


15.3 Multiple linear

4 BARDIA  n  1.9100E+02  1000.  2807
5 OCCWT   n  7.2000E+01  83.00  3265
6 OCCWT_  m  2
:
106 CRBANG p  0.000  315.0  360  24
107 PDOF   p  0.000  345.0  360  23
108 CARANG p  0.000  99.00  360  991
109 VEHOR  p  0.000  90.00  360  995
110 RSTFRT c  3
111 HIC2   d  0.000  1.000
112 estHIC2 e  0.000  0.8455

=================== Constructed variables ===================
113 BARSHP.134 f  0.000  1.000
114 BARSHP.488 f  0.000  1.000
115 BARSHP.EOB f  0.000  1.000
116 BARSHP.EOL f  0.000  1.000
117 BARSHP.FAB f  0.000  1.000
118 BARSHP.FLB f  0.000  1.000
119 BARSHP.GRL f  0.000  1.000
120 BARSHP.IAT f  0.000  1.000
121 BARSHP.LCB f  0.000  1.000
122 BARSHP.LUM f  0.000  1.000
123 BARSHP.MBR f  0.000  1.000
124 BARSHP.OTH f  0.000  1.000
125 BARSHP.PDU f  0.000  1.000
126 BARSHP.POL f  0.000  1.000
127 BARSHP.ROR f  0.000  1.000
128 BARSHP.SGN f  0.000  1.000
129 BARSHP.UNK f  0.000  1.000
130 BARSHP.US1 f  0.000  1.000
131 BARSHP.US2 f  0.000  1.000
132 BARSHP.US3 f  0.000  1.000
133 BARDIA.NA f  0.000  1.000
134 OCCWT.NA f  0.000  1.000
:
178 BX20.NA f  0.000  1.000
179 BX21.NA f  0.000  1.000
180 VEHSPD.NA f  0.000  1.000

Total #cases w/ #missing
#cases miss. D ord. vals #X-var #N-var #F-var #S-var
3310  34  3310  1  48  48  0

#P-var #M-var #B-var #C-var #I-var
6  42  1  12  0

Number of cases used for training: 3276
Number of split variables: 61
Number of dummy variables created: 20
Number of cases excluded due to 0 W or missing D variable: 34
Proportion of ones in HIC2 variable: 0.084554

Missing regressors imputed with means and missing-value indicators added
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE’s for pruned tree: 0.2500

Nodewise interaction tests on all variables
Fraction of cases used for splitting each node: 1.0000
Maximum number of split levels: 13
Minimum node sample size: 63
Minimum number of D=0 and D=1 in each node: 9
150 bootstrap calibration replicates
Scaling for N variables after bootstrap calibration: 1.000

Top-ranked variables and 1-df chi-squared values at root node

<table>
<thead>
<tr>
<th>Rank</th>
<th>Value</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.7106E+03</td>
<td>MODELD</td>
</tr>
<tr>
<td>2</td>
<td>0.6311E+03</td>
<td>COLMEC</td>
</tr>
<tr>
<td>3</td>
<td>0.4562E+03</td>
<td>YEAR</td>
</tr>
<tr>
<td>4</td>
<td>0.4439E+03</td>
<td>RSTFRT</td>
</tr>
<tr>
<td>63</td>
<td>0.3634E+01</td>
<td>OFFSET</td>
</tr>
<tr>
<td>64</td>
<td>0.1315E+01</td>
<td>CRBANG</td>
</tr>
<tr>
<td>65</td>
<td>0.3776E-01</td>
<td>IMPANG</td>
</tr>
<tr>
<td>66</td>
<td>0.6323E-02</td>
<td>BARDIA</td>
</tr>
</tbody>
</table>

Size and CV Loss and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Nodes</th>
<th>Mean Loss</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median Loss</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14</td>
<td>1.798E+08</td>
<td>1.798E+08</td>
<td>1.798E+08</td>
<td>Infinity</td>
<td>1.798E+08</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>1.798E+08</td>
<td>1.798E+08</td>
<td>1.798E+08</td>
<td>Infinity</td>
<td>1.798E+08</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>1.798E+08</td>
<td>1.798E+08</td>
<td>1.798E+08</td>
<td>Infinity</td>
<td>1.798E+08</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>1.798E+08</td>
<td>1.798E+08</td>
<td>1.798E+08</td>
<td>Infinity</td>
<td>1.798E+08</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>1.798E+08</td>
<td>1.798E+08</td>
<td>1.798E+08</td>
<td>Infinity</td>
<td>1.798E+08</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>1.798E+08</td>
<td>1.798E+08</td>
<td>1.798E+08</td>
<td>Infinity</td>
<td>1.798E+08</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>1.798E+08</td>
<td>1.798E+08</td>
<td>1.798E+08</td>
<td>Infinity</td>
<td>1.798E+08</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>1.798E+08</td>
<td>1.798E+08</td>
<td>1.798E+08</td>
<td>5.013E-01</td>
<td>1.798E+08</td>
</tr>
<tr>
<td>9**</td>
<td>1</td>
<td>1.798E+08</td>
<td>1.798E+08</td>
<td>1.798E+08</td>
<td>5.386E-01</td>
<td>1.798E+08</td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 1 terminal node
0-SE tree based on median is marked with + and has 1 terminal node
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
* tree, ** tree, + tree, and ++ tree all the same
Following tree is based on mean CV with naive SE estimate (**)

Structure of final tree. Each terminal node is marked with a T.

D-mean is mean of HIC2 in the node
Cases fit give the number of cases used to fit node
Node deviance is residual deviance divided by residual degrees of freedom

<table>
<thead>
<tr>
<th>Node label</th>
<th>Total cases</th>
<th>Matrix node</th>
<th>Node split</th>
<th>Other variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1T</td>
<td>3276</td>
<td>3276</td>
<td>108</td>
<td>8.455E-02, 5.509E-01</td>
</tr>
</tbody>
</table>

Best split at root node is on COLMEC

Number of terminal nodes of final tree: 1
Total number of nodes of final tree: 1
Best split variable (based on curvature test) at root node is COLMEC

Regression tree:

Node 1: HIC2 proportion of 1s = 0.84554335E-1

***************************************************************
Predictor means below are means of cases with no missing values.
Regression coefficients are computed from the complete cases.

Node 1: Terminal node

Coefficients of logit:

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Coefficient</th>
<th>t-stat</th>
<th>p-value</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-8.116</td>
<td>-0.6531E-01</td>
<td>0.9479</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARDIA</td>
<td>0.8790E-03</td>
<td>0.3432</td>
<td>0.7315</td>
<td>191.0</td>
<td>284.6</td>
<td>1000.</td>
</tr>
<tr>
<td>OCCWT</td>
<td>0.2460</td>
<td>0.5460</td>
<td>0.5851</td>
<td>72.00</td>
<td>77.73</td>
<td>83.00</td>
</tr>
<tr>
<td>HH</td>
<td>0.1349E-02</td>
<td>0.9004</td>
<td>0.3680</td>
<td>58.00</td>
<td>360.7</td>
<td>4321.</td>
</tr>
<tr>
<td>HW</td>
<td>-0.3223E-03</td>
<td>-0.2249</td>
<td>0.8221</td>
<td>163.0</td>
<td>601.1</td>
<td>6355.</td>
</tr>
<tr>
<td>HR</td>
<td>0.2768E-03</td>
<td>0.2094</td>
<td>0.8341</td>
<td>38.00</td>
<td>223.2</td>
<td>2801.</td>
</tr>
<tr>
<td>HS</td>
<td>-0.2249E-02</td>
<td>-0.9436</td>
<td>0.3454</td>
<td>94.00</td>
<td>309.8</td>
<td>3051.</td>
</tr>
<tr>
<td>CD</td>
<td>0.1583E-03</td>
<td>0.8754E-01</td>
<td>0.9303</td>
<td>241.0</td>
<td>529.7</td>
<td>5857.</td>
</tr>
<tr>
<td>CS</td>
<td>-0.1346E-03</td>
<td>-0.7103E-01</td>
<td>0.9434</td>
<td>41.00</td>
<td>295.7</td>
<td>4077.</td>
</tr>
<tr>
<td>AD</td>
<td>0.1287E-03</td>
<td>0.8821E-01</td>
<td>0.9297</td>
<td>10.00</td>
<td>124.4</td>
<td>7625.</td>
</tr>
<tr>
<td>HD</td>
<td>0.3573E-03</td>
<td>0.1616</td>
<td>0.8717</td>
<td>20.00</td>
<td>162.5</td>
<td>1000.</td>
</tr>
<tr>
<td>KD</td>
<td>-0.3990E-03</td>
<td>-0.1847</td>
<td>0.8535</td>
<td>2.000</td>
<td>139.8</td>
<td>315.0</td>
</tr>
<tr>
<td>HB</td>
<td>-0.5234E-01</td>
<td>-0.4176</td>
<td>0.6763</td>
<td>83.00</td>
<td>800.7</td>
<td>1000.</td>
</tr>
<tr>
<td>NB</td>
<td>0.1794</td>
<td>0.5362</td>
<td>0.5919</td>
<td>185.0</td>
<td>917.7</td>
<td>1000.</td>
</tr>
<tr>
<td>CB</td>
<td>0.6487E-01</td>
<td>0.5378</td>
<td>0.5908</td>
<td>200.0</td>
<td>863.0</td>
<td>1000.</td>
</tr>
<tr>
<td>KB</td>
<td>-0.1336</td>
<td>-0.5746</td>
<td>0.5656</td>
<td>-10.00</td>
<td>8.139</td>
<td>1000.</td>
</tr>
<tr>
<td>CLSSPD</td>
<td>0.1903E-01</td>
<td>0.5326</td>
<td>0.5944</td>
<td>19.60</td>
<td>50.37</td>
<td>99.10</td>
</tr>
<tr>
<td>OFFSET</td>
<td>-0.2703E-03</td>
<td>-0.4019</td>
<td>0.6878</td>
<td>-1054.</td>
<td>-24.73</td>
<td>900.0</td>
</tr>
<tr>
<td>IMPPT</td>
<td>-0.1379E-03</td>
<td>-0.1875</td>
<td>0.8513</td>
<td>-690.0</td>
<td>22.60</td>
<td>1739.</td>
</tr>
</tbody>
</table>
15.3 Multiple linear

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>t-value</th>
<th>p-value</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>YEAR</td>
<td>-0.3630E-01</td>
<td>-1.812</td>
<td>0.7001E-01</td>
<td>1972.</td>
<td>2000.</td>
</tr>
<tr>
<td>ENGDSP</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.6000</td>
<td>0.243E+05</td>
</tr>
<tr>
<td>VEHTWT</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>728.0</td>
<td>1758.</td>
</tr>
<tr>
<td>CUBWT</td>
<td>0.1303E-03</td>
<td>0.2514</td>
<td>0.8015</td>
<td>964.0</td>
<td>1704.</td>
</tr>
<tr>
<td>WLBAS</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>299.0</td>
<td>2755.</td>
</tr>
<tr>
<td>VEHLEN</td>
<td>-0.3087E-04</td>
<td>-0.1183</td>
<td>0.9059</td>
<td>547.0</td>
<td>4708.</td>
</tr>
<tr>
<td>VEHWID</td>
<td>-0.2450E-03</td>
<td>-0.4611</td>
<td>0.6412</td>
<td>185.0</td>
<td>1781.</td>
</tr>
<tr>
<td>VEHC</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>109.0</td>
<td>1246.</td>
</tr>
<tr>
<td>BX1</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>481.0</td>
<td>4713.</td>
</tr>
<tr>
<td>BX2</td>
<td>-0.1913E-03</td>
<td>-0.5081</td>
<td>0.6114</td>
<td>260.0</td>
<td>4062.</td>
</tr>
<tr>
<td>BX3</td>
<td>0.2324E-03</td>
<td>0.3746</td>
<td>0.7080</td>
<td>496.0</td>
<td>3628.</td>
</tr>
<tr>
<td>BX4</td>
<td>0.1566E-03</td>
<td>0.3952</td>
<td>0.6927</td>
<td>1715.</td>
<td>9500.</td>
</tr>
<tr>
<td>BX5</td>
<td>-0.3331E-03</td>
<td>-0.3123</td>
<td>0.7549</td>
<td>80.00</td>
<td>3224.</td>
</tr>
<tr>
<td>BX6</td>
<td>0.8214E-03</td>
<td>0.6416</td>
<td>0.5212</td>
<td>339.0</td>
<td>3217.</td>
</tr>
<tr>
<td>BX7</td>
<td>-0.4450E-03</td>
<td>-0.3873</td>
<td>0.6986</td>
<td>153.0</td>
<td>3207.</td>
</tr>
<tr>
<td>BX8</td>
<td>0.1878E-03</td>
<td>0.1459</td>
<td>0.8840</td>
<td>155.0</td>
<td>2151.</td>
</tr>
<tr>
<td>BX9</td>
<td>0.6013E-04</td>
<td>0.3879E-01</td>
<td>0.9691</td>
<td>161.0</td>
<td>2149.</td>
</tr>
<tr>
<td>BX10</td>
<td>0.2516E-03</td>
<td>0.1553</td>
<td>0.8766</td>
<td>167.0</td>
<td>2165.</td>
</tr>
<tr>
<td>BX11</td>
<td>-0.3446E-04</td>
<td>-0.3161E-01</td>
<td>0.9748</td>
<td>25.00</td>
<td>2165.</td>
</tr>
<tr>
<td>BX12</td>
<td>-0.9530E-04</td>
<td>-0.9348E-01</td>
<td>0.9255</td>
<td>214.0</td>
<td>3213.</td>
</tr>
<tr>
<td>BX13</td>
<td>-0.2188E-03</td>
<td>-0.1696</td>
<td>0.8653</td>
<td>80.00</td>
<td>3214.</td>
</tr>
<tr>
<td>BX14</td>
<td>0.6497E-05</td>
<td>0.5840E-01</td>
<td>0.9534</td>
<td>289.0</td>
<td>3561.</td>
</tr>
<tr>
<td>BX15</td>
<td>-0.1788E-03</td>
<td>-0.2931</td>
<td>0.7695</td>
<td>341.0</td>
<td>3552.</td>
</tr>
<tr>
<td>BX16</td>
<td>-0.4620E-04</td>
<td>-0.6165E-01</td>
<td>0.9508</td>
<td>10.00</td>
<td>2772.</td>
</tr>
<tr>
<td>BX17</td>
<td>-0.3180E-03</td>
<td>-0.6487</td>
<td>0.5166</td>
<td>75.00</td>
<td>408.9</td>
</tr>
<tr>
<td>BX18</td>
<td>0.1084E-03</td>
<td>0.2331</td>
<td>0.8157</td>
<td>10.00</td>
<td>468.4</td>
</tr>
<tr>
<td>BX19</td>
<td>-0.2192E-04</td>
<td>-0.2132</td>
<td>0.8312</td>
<td>33.00</td>
<td>4520.</td>
</tr>
<tr>
<td>BX20</td>
<td>-0.7397E-04</td>
<td>-0.2273</td>
<td>0.8202</td>
<td>356.0</td>
<td>4505.</td>
</tr>
<tr>
<td>BX21</td>
<td>0.2424E-04</td>
<td>0.7224E-01</td>
<td>0.9424</td>
<td>54.00</td>
<td>499.6</td>
</tr>
<tr>
<td>VEHSPD</td>
<td>0.1352E-02</td>
<td>0.4181E-01</td>
<td>0.9667</td>
<td>0.3000</td>
<td>50.30</td>
</tr>
<tr>
<td>BARSHP.134</td>
<td>0.9360E-01</td>
<td>0.8664E-01</td>
<td>0.9310</td>
<td>0.000</td>
<td>0.2625E-01</td>
</tr>
<tr>
<td>BARSHP.488</td>
<td>-0.2838</td>
<td>-0.8350E-01</td>
<td>0.9335</td>
<td>0.000</td>
<td>0.6105E-03</td>
</tr>
<tr>
<td>BARSHP.EOB</td>
<td>-0.3344</td>
<td>-0.2943</td>
<td>0.7685</td>
<td>0.000</td>
<td>0.4915E-01</td>
</tr>
<tr>
<td>BARSHP.EOL</td>
<td>-0.4358</td>
<td>-0.1692</td>
<td>0.8657</td>
<td>0.000</td>
<td>0.9158E-03</td>
</tr>
<tr>
<td>BARSHP.FAB</td>
<td>-0.1791</td>
<td>-0.1480</td>
<td>0.8824</td>
<td>0.000</td>
<td>0.1160E-01</td>
</tr>
<tr>
<td>BARSHP.FLB</td>
<td>-0.1395E-01</td>
<td>-0.1368E-01</td>
<td>0.9891</td>
<td>0.000</td>
<td>0.3126</td>
</tr>
<tr>
<td>BARSHP.GRL</td>
<td>-2.054</td>
<td>-1.106</td>
<td>0.2688</td>
<td>0.000</td>
<td>0.2747E-02</td>
</tr>
<tr>
<td>BARSHP.IAT</td>
<td>-2.103</td>
<td>-1.292</td>
<td>0.1966</td>
<td>0.000</td>
<td>0.5800E-02</td>
</tr>
<tr>
<td>BARSHP.LCB</td>
<td>0.1818</td>
<td>0.1816</td>
<td>0.8559</td>
<td>0.000</td>
<td>0.4173</td>
</tr>
<tr>
<td>BARSHP.LUM</td>
<td>1.979</td>
<td>1.070</td>
<td>0.2848</td>
<td>0.000</td>
<td>0.4884E-02</td>
</tr>
<tr>
<td>BARSHP.MBR</td>
<td>-1.862</td>
<td>-0.5974</td>
<td>0.5503</td>
<td>0.000</td>
<td>0.6105E-03</td>
</tr>
<tr>
<td>BARSHP.OTH</td>
<td>-0.2325</td>
<td>-0.1745</td>
<td>0.8615</td>
<td>0.000</td>
<td>0.6716E-02</td>
</tr>
<tr>
<td>BARSHP.PDU</td>
<td>0.6018E-01</td>
<td>0.3244E-01</td>
<td>0.9741</td>
<td>0.000</td>
<td>0.2137E-02</td>
</tr>
<tr>
<td>BARSHP.POL</td>
<td>1.468</td>
<td>1.008</td>
<td>0.3136</td>
<td>0.000</td>
<td>0.1371</td>
</tr>
<tr>
<td>BARSHP.ROH</td>
<td>-0.8234</td>
<td>-0.2690</td>
<td>0.7879</td>
<td>0.000</td>
<td>0.6105E-03</td>
</tr>
<tr>
<td>BARSHP.SGN</td>
<td>-2.242</td>
<td>-0.9906</td>
<td>0.3220</td>
<td>0.000</td>
<td>0.1526E-02</td>
</tr>
</tbody>
</table>

Wei-Yin Loh 282  GUIDE manual
<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient 1</th>
<th>Coefficient 2</th>
<th>Coefficient 3</th>
<th>p-value</th>
<th>SE</th>
<th>z-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BARSHP_UNK</td>
<td>-0.3980</td>
<td>-0.1177</td>
<td>0.9063</td>
<td>0.000</td>
<td>0.1221E-02</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BARSHP_US1</td>
<td>-0.2868</td>
<td>-0.1619</td>
<td>0.8714</td>
<td>0.000</td>
<td>0.2442E-02</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BARSHP_US2</td>
<td>0.6299E-01</td>
<td>0.4168E-01</td>
<td>0.9668</td>
<td>0.000</td>
<td>0.4579E-02</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BARSHP_US3</td>
<td>-0.5868E-01</td>
<td>-0.4326E-01</td>
<td>0.9655</td>
<td>0.000</td>
<td>0.6105E-02</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BARDIA_NA</td>
<td>0.8465</td>
<td>0.8689</td>
<td>0.3849</td>
<td>0.000</td>
<td>0.8568</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>OCCWT_NA</td>
<td>0.3094</td>
<td>0.1923</td>
<td>0.8475</td>
<td>0.000</td>
<td>0.9966</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>HH_NA</td>
<td>0.9053</td>
<td>0.8528</td>
<td>0.3938</td>
<td>0.000</td>
<td>0.4579E-01</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>HW_NA</td>
<td>-0.5159</td>
<td>-0.3919</td>
<td>0.6951</td>
<td>0.000</td>
<td>0.3816E-01</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>HR_NA</td>
<td>-0.1275</td>
<td>-0.7099E-01</td>
<td>0.9434</td>
<td>0.000</td>
<td>0.5830E-01</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>HS_NA</td>
<td>0.8096E-01</td>
<td>0.6311E-01</td>
<td>0.9497</td>
<td>0.000</td>
<td>0.6105E-02</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>CD_NA</td>
<td>0.1951</td>
<td>0.4839</td>
<td>0.6285</td>
<td>0.000</td>
<td>0.1300</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>CS_NA</td>
<td>-0.8660</td>
<td>-0.8299</td>
<td>0.4067</td>
<td>0.000</td>
<td>0.4518E-01</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>AD_NA</td>
<td>-1.063</td>
<td>-0.6113</td>
<td>0.5410</td>
<td>0.000</td>
<td>0.5830E-01</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>HD_NA</td>
<td>1.360</td>
<td>0.7498</td>
<td>0.4535</td>
<td>0.000</td>
<td>0.6044E-01</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>KD_NA</td>
<td>0.1147</td>
<td>0.1530</td>
<td>0.8784</td>
<td>0.000</td>
<td>0.4976E-01</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>HB_NA</td>
<td>31.23</td>
<td>0.4179</td>
<td>0.6761</td>
<td>0.000</td>
<td>0.9936</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>NB_NA</td>
<td>-133.6</td>
<td>-0.5371</td>
<td>0.5912</td>
<td>0.000</td>
<td>0.9945</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>CB_NA</td>
<td>-17.77</td>
<td>-0.6137</td>
<td>0.5395</td>
<td>0.000</td>
<td>0.9939</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>KB_NA</td>
<td>0.9033</td>
<td>0.4730</td>
<td>0.6362</td>
<td>0.000</td>
<td>0.4014</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>CLSSPD_NA</td>
<td>18.63</td>
<td>0.4698</td>
<td>0.6386</td>
<td>0.000</td>
<td>0.6105E-03</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>OFFSET_NA</td>
<td>-0.1018</td>
<td>-0.4149</td>
<td>0.6783</td>
<td>0.000</td>
<td>0.1401</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>IMPNNT_NA</td>
<td>0.6058E-01</td>
<td>0.1270</td>
<td>0.8989</td>
<td>0.000</td>
<td>0.5168</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>YEAR_NA</td>
<td>1.150</td>
<td>0.5182</td>
<td>0.6044</td>
<td>0.000</td>
<td>0.1221E-02</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>ENGDSP_NA</td>
<td>0.8412E-01</td>
<td>0.1364</td>
<td>0.8915</td>
<td>0.000</td>
<td>0.1709E-01</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>VEHTWT_NA</td>
<td>0.1163</td>
<td>0.6647E-01</td>
<td>0.9470</td>
<td>0.000</td>
<td>0.1832E-02</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>CURBWT_NA</td>
<td>-0.9897E-01</td>
<td>-0.3568</td>
<td>0.7213</td>
<td>0.000</td>
<td>0.8712</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>WHLBAS_NA</td>
<td>0.3689</td>
<td>0.4035</td>
<td>0.6866</td>
<td>0.000</td>
<td>0.1282E-01</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>VEHELSEN_NA</td>
<td>-0.2320</td>
<td>-0.1352</td>
<td>0.8925</td>
<td>0.000</td>
<td>0.4579E-02</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>VEHWID_NA</td>
<td>-0.2375</td>
<td>-0.5213</td>
<td>0.6022</td>
<td>0.000</td>
<td>0.3388E-01</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>VEHC.G_NA</td>
<td>-0.1312</td>
<td>-0.2109</td>
<td>0.8330</td>
<td>0.000</td>
<td>0.5098E-01</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BX1.NA</td>
<td>-0.7540</td>
<td>-0.4347</td>
<td>0.6638</td>
<td>0.000</td>
<td>0.1624</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BX2.NA</td>
<td>0.7244</td>
<td>0.2546</td>
<td>0.7990</td>
<td>0.000</td>
<td>0.2900</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BX10.NA</td>
<td>25.47</td>
<td>0.4125</td>
<td>0.6800</td>
<td>0.000</td>
<td>0.2888</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BX11.NA</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.2891</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BX12.NA</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.2888</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BX13.NA</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.2888</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BX14.NA</td>
<td>0.1146</td>
<td>0.2853E-01</td>
<td>0.9772</td>
<td>0.000</td>
<td>0.2891</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BX15.NA</td>
<td>0.8082</td>
<td>0.3816</td>
<td>0.7028</td>
<td>0.000</td>
<td>0.2903</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BX16.NA</td>
<td>-0.8212E-01</td>
<td>-0.3352E-01</td>
<td>0.9733</td>
<td>0.000</td>
<td>0.2845</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BX17.NA</td>
<td>-3.485</td>
<td>-0.5043</td>
<td>0.6141</td>
<td>0.000</td>
<td>0.2894</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BX18.NA</td>
<td>0.6419</td>
<td>0.4455</td>
<td>0.6560</td>
<td>0.000</td>
<td>0.2845</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BX19.NA</td>
<td>0.2841</td>
<td>0.6351E-01</td>
<td>0.9494</td>
<td>0.000</td>
<td>0.1648</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BX20.NA</td>
<td>0.2063</td>
<td>0.4209E-01</td>
<td>0.9664</td>
<td>0.000</td>
<td>0.1645</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BX21.NA</td>
<td>0.4847</td>
<td>0.4262</td>
<td>0.6700</td>
<td>0.000</td>
<td>0.2943</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BX3.NA</td>
<td>-0.2333E-01</td>
<td>-0.1048E-01</td>
<td>0.9916</td>
<td>0.000</td>
<td>0.2903</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BX4.NA</td>
<td>0.6271</td>
<td>0.1868</td>
<td>0.8518</td>
<td>0.000</td>
<td>0.2897</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Variable</td>
<td>Coefficient</td>
<td>Standard Error</td>
<td>z Value</td>
<td>p Value</td>
<td>Odd Ratio</td>
<td>95% CI Lower</td>
<td>95% CI Upper</td>
</tr>
<tr>
<td>----------</td>
<td>-------------</td>
<td>----------------</td>
<td>---------</td>
<td>---------</td>
<td>-----------</td>
<td>--------------</td>
<td>--------------</td>
</tr>
<tr>
<td>BX5.NA</td>
<td>-0.7012</td>
<td>0.1743</td>
<td>0.8616</td>
<td>0.000</td>
<td>0.2894</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BX6.NA</td>
<td>-1.194</td>
<td>0.3422</td>
<td>0.7322</td>
<td>0.000</td>
<td>0.2894</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BX7.NA</td>
<td>-23.38</td>
<td>0.3854</td>
<td>0.6999</td>
<td>0.000</td>
<td>0.2891</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BX8.NA</td>
<td>0.3424</td>
<td>0.8482E-01</td>
<td>0.9324</td>
<td>0.000</td>
<td>0.2891</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BX9.NA</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.2891</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>VEHSPD.NA</td>
<td>-0.3514</td>
<td>-0.1538</td>
<td>0.8777</td>
<td>0.000</td>
<td>0.1632E-02</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

Proportion of ones in variable HIC2 = 0.845543E-1

--------------------
Observed and fitted values are stored in logitm.fit
LaTeX code for tree is in logitm.tex
R code is stored in logitm.r
16 Importance scoring

When there are numerous predictor variables, it may be useful to rank them in order of their “importance”. GUIDE has a facility to do this. In addition, it provides thresholds for grouping the variables by their importance—see Loh and Zhou (2021).

16.1 Classification: RHC data

We show here how to obtain the importance scores for predicting \texttt{swang1}, the variable that takes values \texttt{RHC} and \texttt{NoRHC}; see Section 4.

16.1.1 Input file creation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: \texttt{imp.in}
   Input 1 for model fitting, 2 for importance or DIF scoring, 3 for data conversion ([1:3], <cr>=1): 2
   Name of batch output file: \texttt{imp.out}
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1):
   Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
   Input name of data description file (max 100 characters); enclose with matching quotes if it has spaces: \texttt{rhcdsc1.txt}
   Reading data description file ...
   Training sample file: \texttt{rhcdata.txt}
   Missing value code: \texttt{NA}
   Records in data file start on line 2
   23 N variables changed to S
   D variable is \texttt{swang1}
   Reading data file ...
   Number of records in data file: 5735
   Length of longest entry in data file: 19
   Checking for missing values ...
   Finished checking
   Missing values found among categorical variables
   Separate categories will be created for missing categorical variables
   Missing values found among non-categorical variables
   Recoding D values to integers
   Finished recoding
   Number of classes: 2
   Assigning integer codes to values of 30 categorical variables
   Finished assigning codes to 10 categorical variables
Finished assigning codes to 20 categorical variables
Finished assigning codes to 30 categorical variables
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...
Finished processing 5000 of 5735 observations
Data checks complete
Creating missing value indicators ...
Rereading data ...
Class #Cases Proportion
NoRHC 3551 0.61918047
RHC 2184 0.38081953
Total #cases w/ #missing
#cases miss. D ord. vals #X-var #N-var #F-var #S-var
5735 0 5157 10 0 0 23
#P-var #M-var #B-var #C-var #I-var
0 0 0 30 0
Number of cases used for training: 5735
Number of split variables: 53
Number of cases excluded due to 0 W or missing D variable: 0
Finished reading data file
Choose 1 for estimated priors, 2 for equal priors, 3 to input priors from a file
Input 1, 2, or 3 ([1:3], <cr>=1):
Choose 1 for unit misclassification costs, 2 to input costs from a file
Input 1 or 2 ([1:2], <cr>=1):
Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):
Input file name to store LaTeX code (use .tex as suffix): imp.tex
You can store the variables and/or values used to split and fit in a file
Choose 1 to skip this step, 2 to store split and fit variables,
3 to store split variables and their values
Input your choice ([1:3], <cr>=1):
You can create a description file with the selected variables included or excluded
Input 2 to create such a file, 1 otherwise ([1:2], <cr>=1):
You can also output the importance scores and variable names to a file
Input 1 to create such a file, 2 otherwise ([1:2], <cr>=1):
Input file name: imp.scr
Input file is created!
Run GUIDE with the command: guide < imp.in

16.1.2 Contents of imp.out
The most interesting part of the output file is at the end, as shown below.

Scaled importance scores of predictor variables

<table>
<thead>
<tr>
<th>Score</th>
<th>Rank</th>
<th>Variable</th>
</tr>
</thead>
</table>

Wei-Yin Loh 286 GUIDE manual
<table>
<thead>
<tr>
<th>Score</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.375E+01</td>
<td>aps1</td>
</tr>
<tr>
<td>2.319E+01</td>
<td>cat1</td>
</tr>
<tr>
<td>2.154E+01</td>
<td>creal</td>
</tr>
<tr>
<td>2.091E+01</td>
<td>pafi1</td>
</tr>
<tr>
<td>1.869E+01</td>
<td>meanbp1</td>
</tr>
<tr>
<td>1.249E+01</td>
<td>neuro</td>
</tr>
<tr>
<td>1.139E+01</td>
<td>alb1</td>
</tr>
<tr>
<td>1.043E+01</td>
<td>card</td>
</tr>
<tr>
<td>1.043E+01</td>
<td>cat2</td>
</tr>
<tr>
<td>1.032E+01</td>
<td>hema1</td>
</tr>
<tr>
<td>9.504E+00</td>
<td>wtkilo1</td>
</tr>
<tr>
<td>8.182E+00</td>
<td>adld3p</td>
</tr>
<tr>
<td>8.131E+00</td>
<td>seps</td>
</tr>
<tr>
<td>6.673E+00</td>
<td>dnr1</td>
</tr>
<tr>
<td>6.567E+00</td>
<td>bili1</td>
</tr>
<tr>
<td>6.348E+00</td>
<td>resp</td>
</tr>
<tr>
<td>5.646E+00</td>
<td>paco21</td>
</tr>
<tr>
<td>5.555E+00</td>
<td>surv2md1</td>
</tr>
<tr>
<td>4.160E+00</td>
<td>transhx</td>
</tr>
<tr>
<td>3.957E+00</td>
<td>chrpulhx</td>
</tr>
<tr>
<td>3.889E+00</td>
<td>hrt1</td>
</tr>
<tr>
<td>3.821E+00</td>
<td>resp1</td>
</tr>
<tr>
<td>3.571E+00</td>
<td>ph1</td>
</tr>
<tr>
<td>3.423E+00</td>
<td>ninsclas</td>
</tr>
<tr>
<td>3.183E+00</td>
<td>dementhx</td>
</tr>
<tr>
<td>2.400E+00</td>
<td>das2d3pc</td>
</tr>
<tr>
<td>2.316E+00</td>
<td>psychhx</td>
</tr>
<tr>
<td>2.118E+00</td>
<td>gastr</td>
</tr>
<tr>
<td>2.083E+00</td>
<td>renal</td>
</tr>
<tr>
<td>1.799E+00</td>
<td>cardiohx</td>
</tr>
<tr>
<td>1.744E+00</td>
<td>income</td>
</tr>
<tr>
<td>1.454E+00</td>
<td>urin1</td>
</tr>
<tr>
<td>1.336E+00</td>
<td>trauma</td>
</tr>
<tr>
<td>1.224E+00</td>
<td>age</td>
</tr>
<tr>
<td>1.168E+00</td>
<td>sex</td>
</tr>
<tr>
<td>1.166E+00</td>
<td>edu</td>
</tr>
<tr>
<td>1.165E+00</td>
<td>sod1</td>
</tr>
<tr>
<td>1.038E+00</td>
<td>wblc1</td>
</tr>
<tr>
<td>9.625E-01</td>
<td>immunhx</td>
</tr>
<tr>
<td>8.970E-01</td>
<td>malighx</td>
</tr>
<tr>
<td>8.870E-01</td>
<td>ca</td>
</tr>
<tr>
<td>8.316E-01</td>
<td>scoma1</td>
</tr>
</tbody>
</table>
The variables, sorted according to their importance scores, are divided into 5 groups:

A. Scores above 99% threshold

B. Scores above 95% threshold and below 99% threshold

C. Scores above 90% threshold and below 95% threshold

D. Scores above 80% threshold and below 90% threshold

E. below 80% threshold

The groups and thresholds have the following interpretation. Let \( H_0 \) denote the null hypothesis \( H_0 \) that the dependent variable is independent of the predictor variables (it is not assumed that the predictor variables are independent of each other). If \( H_0 \) is true, there is a 0.01, 0.05, 0.10, and 0.20 probability that one or more predictor variables falls into groups A, \( A \cup B \), \( A \cup B \cup C \), \( A \cup B \cup C \cup D \), respectively. The importance scores are normalized so that the 95% threshold is 1.0.

The file \texttt{imp.scr} lists the group memberships, scores and variables names:
### 16.1 Classification: RHC data

<table>
<thead>
<tr>
<th>Rank</th>
<th>Type</th>
<th>Score</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>2.375E+01</td>
<td>aps1</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>2.319E+01</td>
<td>cat1</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>2.154E+01</td>
<td>crea1</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>2.091E+01</td>
<td>pafi1</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>1.869E+01</td>
<td>meanbp1</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>1.249E+01</td>
<td>neuro</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>1.139E+01</td>
<td>alb1</td>
</tr>
<tr>
<td>8</td>
<td>A</td>
<td>1.043E+01</td>
<td>card</td>
</tr>
<tr>
<td>9</td>
<td>A</td>
<td>1.043E+01</td>
<td>cat2</td>
</tr>
<tr>
<td>10</td>
<td>A</td>
<td>1.032E+01</td>
<td>hemal</td>
</tr>
<tr>
<td>11</td>
<td>A</td>
<td>9.504E+00</td>
<td>wtkilo1</td>
</tr>
<tr>
<td>12</td>
<td>A</td>
<td>8.182E+00</td>
<td>add3p</td>
</tr>
<tr>
<td>13</td>
<td>A</td>
<td>8.131E+00</td>
<td>seps</td>
</tr>
<tr>
<td>14</td>
<td>A</td>
<td>6.673E+00</td>
<td>drn1</td>
</tr>
<tr>
<td>15</td>
<td>A</td>
<td>6.567E+00</td>
<td>bili1</td>
</tr>
<tr>
<td>16</td>
<td>A</td>
<td>6.348E+00</td>
<td>resp</td>
</tr>
<tr>
<td>17</td>
<td>A</td>
<td>5.646E+00</td>
<td>pac21</td>
</tr>
<tr>
<td>18</td>
<td>A</td>
<td>5.555E+00</td>
<td>surv2md1</td>
</tr>
<tr>
<td>19</td>
<td>A</td>
<td>4.160E+00</td>
<td>transh1</td>
</tr>
<tr>
<td>20</td>
<td>A</td>
<td>3.957E+00</td>
<td>chrpulhx</td>
</tr>
<tr>
<td>21</td>
<td>A</td>
<td>3.889E+00</td>
<td>hrt1</td>
</tr>
<tr>
<td>22</td>
<td>A</td>
<td>3.821E+00</td>
<td>rep1</td>
</tr>
<tr>
<td>23</td>
<td>A</td>
<td>3.571E+00</td>
<td>phi</td>
</tr>
<tr>
<td>24</td>
<td>A</td>
<td>3.423E+00</td>
<td>ninsclas</td>
</tr>
<tr>
<td>25</td>
<td>A</td>
<td>3.183E+00</td>
<td>demethx</td>
</tr>
<tr>
<td>26</td>
<td>A</td>
<td>2.400E+00</td>
<td>das2d3pc</td>
</tr>
<tr>
<td>27</td>
<td>A</td>
<td>2.316E+00</td>
<td>psychhx</td>
</tr>
<tr>
<td>28</td>
<td>A</td>
<td>2.118E+00</td>
<td>gastr</td>
</tr>
<tr>
<td>29</td>
<td>A</td>
<td>2.083E+00</td>
<td>renal</td>
</tr>
<tr>
<td>30</td>
<td>A</td>
<td>1.799E+00</td>
<td>cardiohx</td>
</tr>
<tr>
<td>31</td>
<td>A</td>
<td>1.744E+00</td>
<td>income</td>
</tr>
<tr>
<td>32</td>
<td>A</td>
<td>1.454E+00</td>
<td>urin1</td>
</tr>
<tr>
<td>33</td>
<td>B</td>
<td>1.336E+00</td>
<td>trauma</td>
</tr>
<tr>
<td>34</td>
<td>B</td>
<td>1.224E+00</td>
<td>age</td>
</tr>
<tr>
<td>35</td>
<td>B</td>
<td>1.168E+00</td>
<td>sex</td>
</tr>
<tr>
<td>36</td>
<td>B</td>
<td>1.166E+00</td>
<td>edu</td>
</tr>
<tr>
<td>37</td>
<td>B</td>
<td>1.165E+00</td>
<td>sod1</td>
</tr>
<tr>
<td>38</td>
<td>B</td>
<td>1.038E+00</td>
<td>wbcl1</td>
</tr>
<tr>
<td>39</td>
<td>C</td>
<td>9.625E-01</td>
<td>immunhx</td>
</tr>
<tr>
<td>40</td>
<td>C</td>
<td>8.970E-01</td>
<td>malighx</td>
</tr>
<tr>
<td>41</td>
<td>D</td>
<td>8.870E-01</td>
<td>ca</td>
</tr>
<tr>
<td>42</td>
<td>E</td>
<td>8.316E-01</td>
<td>scomal</td>
</tr>
<tr>
<td>43</td>
<td>E</td>
<td>8.198E-01</td>
<td>amihx</td>
</tr>
<tr>
<td>44</td>
<td>E</td>
<td>6.724E-01</td>
<td>chfhx</td>
</tr>
<tr>
<td>45</td>
<td>E</td>
<td>6.284E-01</td>
<td>gibledhx</td>
</tr>
</tbody>
</table>
Figure 35 shows a barplot of the scores. It is produced by the following R code.

```r
par(las=1,mar=c(5,12,4,2),cex.axis=0.8)
leg.col <- c("red","orange","yellow","green","white")
leg.txt <- c("A (99%)","B (95%)","C (90%)","D (80%)","E (< 80%)")
x <- read.table("imp.scr",header=TRUE)
score <- x$Score
vars <- x$Variable
type <- x$Type
barcol <- rep("white",n)
letrs <- c("A","B","C","D","E")
for(i in 1:4){
  barcol[type == letrs[i]] <- leg.col[i]
}
n <- nrow(x)
barplot(rev(score[1:n]),names.arg=rev(vars[1:n]),col=rev(barcol[1:n]),horiz=TRUE,
  xlab="GUIDE importance scores")
legend("bottomright",legend=leg.txt,fill=leg.col)
```

Figure 36 shows the classification tree from `imp.tex` that produced the scores. It is an unpruned tree with four levels of splits.
Figure 35: Scores of important variables for predicting swang1
Figure 36: GUIDE v.41.1 importance scoring classification tree for predicting \textit{swang1} using estimated priors and unit misclassification costs. At each split, an observation goes to the left branch if and only if the condition is satisfied. Symbol ‘$\leq_*$’ stands for ‘$\leq$ or missing’. $S_1 = \{\text{CHF, MOSF w/Sepsis}\}$. $S_2 = \{\text{No insurance, Private, Private & Medicare}\}$. Predicted classes and sample sizes (in italics) printed below terminal nodes; class sample proportion for $\textit{swang1} = \text{RHC}$ beside nodes. Second best split variable at root node is $\text{aps1}$.
16.2 Censored response with R variable

Following is the corresponding scoring procedure for a censored response with a treatment (R) variable (swang1). The R variable is not given a score because it acts as a linear predictor in the nodes of the tree.

16.2.1 Input file creation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: imp_surv.in
   Input 1 for model fitting, 2 for importance or DIF scoring,
   3 for data conversion ([1:3], <cr>=1): 2
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1): 2
   Choose type of regression model:
      1=linear, 2=quantile, 3=Poisson, 4=censored response,
      5=multiresponse or itemresponse, 6=longitudinal data (with T variables).
   Input choice ([1:6], <cr>=1): 4
   Input 1 for proportional hazards, 2 for restricted mean event time ([1:2], <cr>=1): 1
   Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
   Input name of data description file (max 100 characters); enclose with matching quotes if it has spaces: rhcdsc3.txt
   Reading data description file ...
   Training sample file: rhcdata.txt
   Missing value code: NA
   Records in data file start on line 2
   R variable present
   23 N variables changed to S
   Warning: model changed to linear in treatment
   D variable is death
   Reading data file ...
   Number of records in data file: 5735
   Length of longest entry in data file: 19
   Checking for missing values ...
   Finished checking
   Missing values found among categorical variables
   Separate categories will be created for missing categorical variables
   Missing values found among non-categorical variables
   Assigning integer codes to values of 31 categorical variables
   Finished assigning codes to 10 categorical variables
   Finished assigning codes to 20 categorical variables
   Finished assigning codes to 30 categorical variables
   Treatment (R) variable is swang1 with values "NoRHC" and "RHC"
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...
Finished processing 5000 of 5735 observations
Data checks complete
Smallest uncensored survtime: 2.0000
Number of complete cases excluding censored T < smallest uncensored T: 5735
Number of cases used to compute baseline hazard: 5735
Number of cases with D=1 and T >= smallest uncensored: 3722
GUIDE will try to create the variables in the description file.
If it is unsuccessful, please create the columns yourself ...
Number of dummy variables created: 1
Input 1 if randomized trial, 2 if observational study: ([1:2], <cr>=1): 2
Choose a subgroup identification method:
1 = Prognostic priority (Gs)
2 = Predictive priority (Gi)
Input your choice: ([1:2], <cr>=2):
Creating dummy variables ...
Creating missing value indicators ...
Rereading data ...
Largest uncensored and censored survtime by swang1
"NoRHC" 1867.0000 1243.0000
"RHC" 1943.0000 1351.0000
Proportion of training sample for each level of swang1
"NoRHC" 0.6192
"RHC" 0.3808
Total #cases w/ #missing
#cases miss. D ord. vals #X-var #N-var #F-var #S-var
5735 0 5157 8 0 0 23
#P-var #M-var #B-var #C-var #I-var #R-var
0 0 0 30 0 1
Survival time variable in column: 64
Event indicator variable in column: 9
Proportion uncensored among nonmissing T and D variables: .649
Number of cases used for training: 5735
Number of split variables: 53
Number of dummy variables created: 1
Number of cases excluded due to 0 W or missing D, T or R variables: 0
Finished reading data file
Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):
Input file name to store LaTeX code (use .tex as suffix): imp_surv.tex
You can store the variables and/or values used to split and fit in a file
Choose 1 to skip this step, 2 to store split and fit variables,
3 to store split variables and their values
Input your choice ([1:3], <cr>=1):
You can create a description file with the selected variables included or excluded
16.2  Censored response with R variable

Input 2 to create such a file, 1 otherwise ([1:2], `<cr>`=1):
You can also output the importance scores and variable names to a file
Input 1 to create such a file, 2 otherwise ([1:2], `<cr>`=1):
Input file name: imp_surv.scr
Input file is created!
Run GUIDE with the command: guide < imp_surv.in

16.2.2  Partial contents of imp_surv.out

The output shows that there is only one important variable.

Scaled importance scores of predictor variables
(F, I and R variables are excluded)
Score  Rank  Variable
1.055E+00 1.00  dnr1
--------- variables above this line have scores above 95\% threshold (B) ------
--------- variables above this line have scores above 90\% threshold (C) ------
9.446E-01 2.00  ph1
8.009E-01 3.00  chrpulhx
--------- variables above this line have scores above 80\% threshold (D) ------
7.860E-01 4.00  resp1
7.851E-01 5.00  paco21
4.947E-01 6.00  liverhx
4.508E-01 7.00  pot1
4.357E-01 8.00  gastr
4.303E-01 9.00  cat2
4.009E-01 10.00 gibledhx
3.967E-01 11.00 age
3.619E-01 12.00 pafi1
3.456E-01 13.00 aps1
3.229E-01 14.00 malighx
3.161E-01 15.00 amihx
2.988E-01 16.00 hrt1
2.856E-01 17.00 surv2md1
2.689E-01 18.00 ninsclas
2.493E-01 19.00 das2d3pc
2.441E-01 20.00 edu
2.440E-01 21.00 meanbp1
2.180E-01 22.00 income
2.000E-01 23.00 scoma1
1.802E-01 24.00 ortho
1.753E-01 25.00 creal
1.736E-01 26.00 templ
1.708E-01 27.00 hemal
1.603E-01 28.00 ca
17 CAUSAL INference

Propensity score matching is often used in causal inference to estimate average treatment effects. Given a treatment variable $Z$ taking values $0$ (no treatment) and $1$ (treatment), the propensity score for a subject with covariate $X = x$ is $\pi(x) = P(Z = 1 \mid X = x)$. If $n$ denotes the sample size and $Y_i$ the response of the
ith subject, the average treatment effect may be estimated by the Horvitz-Thompson estimate (HT)

\[ n^{-1} \sum_{i=1}^{n} \left\{ \frac{Z_i Y_i}{\hat{\pi}(X_i)} - \frac{(1 - Z_i)Y_i}{1 - \hat{\pi}(X_i)} \right\} \]

or the Hájek inverse probability estimate (IPW)

\[ \frac{\sum_i Z_i Y_i / \hat{\pi}(X_i)}{\sum_i Z_i / \hat{\pi}(X_i)} - \frac{\sum_i (1 - Z_i) Y_i / (1 - \hat{\pi}(X_i))}{\sum_i (1 - Z_i) / (1 - \hat{\pi}(X_i))} \]

where \( \hat{\pi}(x) \) is an estimate of \( \pi(x) \). Clearly, \( \hat{\pi}(x) \) cannot be 0 or 1.

The propensity scores are traditionally estimated by logistic regression, but this approach has difficulties if there are missing values in the covariates or if the number of covariates is large. Random forest has been used, but the version implemented in R is not applicable to data with missing values. Even when there are no missing values, the propensity score estimates from logistic regression and random forest are not easy to interpret.

A classification tree for predicting \( Z \) is much more interpretable than a forest, but one or more terminal nodes may be pure (i.e., all \( Z_i = 0 \) or all \( Z_i = 1 \)), resulting in \( \hat{\pi}(x_i) = 0 \) or 1. To avoid this, GUIDE has a “propensity score” option that disallows such splits. Specifically, it only allows splits that yield in each subnode at least \( m \) observations each of \( Z = 0 \) and \( Z = 1 \). The value of \( m \) is a positive integer that may be specified by the user. If a GUIDE piecewise-constant model is used to estimate the propensity scores, the HT and IPW estimates are identical and reduce to the sample size weighted estimate

\[ n^{-1} \sum_t n_t \hat{\beta}_t, \]

where the sum is over the terminal nodes and \( n_t \) and \( \hat{\beta}_t \) are the sample size and estimated treatment effect in node \( t \).

We demonstrate the propensity score feature with the RHC data. Doctors believe that direct measurement of cardiac function by right heart catheterization for some critically ill patients yields better outcomes. The benefit of RHC has not been demonstrated in a randomized clinical trial due to ethical concerns. In observational studies, the relative risk of death was found to be higher in the elderly and in patients with acute myocardial infarction who received RHC. In such studies, the decision to use RHC is at the discretion of the physician. Therefore treatment assignment is confounded with patient factors that are also related to outcomes, e.g., patients with low blood pressure are more likely to get RHC, and such patients are also more likely to die. The data consist of observations on more than 60 variables for 5735 patients from 5 medical centers over 5 years (Connors et al., 1996). The treatment variable is \texttt{swang1} (RHC or NoRHC), and the response variables are \texttt{dth30} (1=death within 30 days, 0=survived more than 30 days) and \texttt{death} (1=eventual death, 0=censored).
17.1 Input file creation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: prop30.in
   Input 1 for model fitting, 2 for importance or DIF scoring,
   3 for data conversion ([1:3], <cr>=1)
   Name of batch output file: prop30.out
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1)
   Input 1 for classification, 2 for regression, 3 for propensity score grouping
   Input your choice ([1:3], <cr>=1): 3
   Input name of data description file (max 100 characters);
   enclose with matching quotes if it has spaces: rhcdsc4.txt
   Reading data description file ...
   Training sample file: rhcdata.txt
   Missing value code: NA
   Records in data file start on line 2
   R variable present
   35 N variables changed to S
   Warning: model changed to linear in treatment
   D variable is dth30
   Reading data file ...
   Number of records in data file: 5735
   Length of longest entry in data file: 19
   Checking for missing values ...
   Finished checking
   Missing values found among categorical variables
   Separate categories will be created for missing categorical variables
   Missing values found among non-categorical variables
   Recoding D values to integers
   Finished recoding
   Number of classes: 2
   Assigning integer codes to values of 18 categorical variables
   Finished assigning codes to 10 categorical variables
   Re-checking data ...
   Allocating missing value information ...
   Assigning codes to missing values, if any ...
   Finished processing 5000 of 5735 observations
   Data checks complete
   Creating missing value indicators ...

The data and description files are rhcdata.txt and rhcdsc4.txt. In the latter, the variable swang1 is designated as r, dth30 as d, and death as x.
## Contents of prop30.out

Rereading data ...

<table>
<thead>
<tr>
<th>Treatment</th>
<th>#Cases</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoRHC</td>
<td>3551</td>
<td>0.61918047</td>
</tr>
<tr>
<td>RHC</td>
<td>2184</td>
<td>0.38081953</td>
</tr>
</tbody>
</table>

| Total #cases w/ #missing #cases miss. D ord. vals #X-var #N-var #F-var #S-var |
|-----------------|---------|---------------|
| 5735            | 5157    | 9             |

<table>
<thead>
<tr>
<th>#P-var #M-var #B-var #C-var #I-var</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>18</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

Number of cases used for training: 5735
Number of split variables: 53
Number of cases excluded due to 0 W or missing D variable: 0
Finishing reading data file

Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):
Input file name to store LaTeX code (use .tex as suffix): prop30.tex
You can store the variables and/or values used to split and fit in a file
Choose 1 to skip this step, 2 to store split and fit variables, 3 to store split variables and their values
Input your choice ([1:3], <cr>=1):
Input 2 to save fitted values and node IDs, 1 otherwise ([1:2], <cr>=2):
Input name of file to store node ID and fitted value of each case: prop30.fit
Input 2 to write R function for predicting new cases, 1 otherwise ([1:2], <cr>=2):
Input file name: prop30.r
Input rank of top variable to split root node ([1:53], <cr>=1):
Input file is created!

Run GUIDE with the command: `guide < prop30.in`

### 17.2 Contents of prop30.out

Propensity score grouping and estimation of causal effects
Pruning by cross-validation
Data description file: rhcdsc4.txt
Training sample file: rhcdata.txt
Missing value code: NA
Records in data file start on line 2
R variable present
35 N variables changed to S
Warning: model changed to linear in treatment
D variable is dth30
Number of records in data file: 5735
Length of longest entry in data file: 19
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Number of classes: 2
17.2 Contents of prop30.out

<table>
<thead>
<tr>
<th>Treatment</th>
<th>#Cases</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoRHC</td>
<td>3551</td>
<td>0.61918047</td>
</tr>
<tr>
<td>RHC</td>
<td>2184</td>
<td>0.38081953</td>
</tr>
</tbody>
</table>

Summary information for training sample of size 5735

d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators),
s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
m=missing-value flag variable, p=periodic variable, w=weight

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>#Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>cat1 c</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>cat2 c</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>ca c</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>cardiohx s</td>
<td>0.000</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>das2d3pc s</td>
<td>11.00</td>
<td>33.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>dth30 d</td>
<td>0.000</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>aps1 s</td>
<td>3.000</td>
<td>147.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>scoma1 s</td>
<td>0.000</td>
<td>100.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>meanbp1 s</td>
<td>10.00</td>
<td>259.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>wblc1 s</td>
<td>0.000</td>
<td>192.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>hrt1 s</td>
<td>8.000</td>
<td>250.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>resp1 s</td>
<td>2.000</td>
<td>100.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>temp1 s</td>
<td>27.00</td>
<td>43.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>pafi1 s</td>
<td>11.60</td>
<td>937.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>alb1 s</td>
<td>0.3000</td>
<td>29.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>hema1 s</td>
<td>2.000</td>
<td>66.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>bili1 s</td>
<td>0.9999E-01</td>
<td>58.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>crea1 s</td>
<td>0.9999E-01</td>
<td>25.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>sod1 s</td>
<td>101.0</td>
<td>259.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>pot1 s</td>
<td>1.100</td>
<td>11.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>paco21 s</td>
<td>1.000</td>
<td>156.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>ph1 s</td>
<td>6.579</td>
<td>7.770</td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>swang1 r</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>wtkilo1 s</td>
<td>19.50</td>
<td>244.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>ortho c</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>addd3p s</td>
<td>0.000</td>
<td>7.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>urin1 s</td>
<td>0.000</td>
<td>9000.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>race c</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>income c</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total #cases w/</th>
<th>#cases miss. D ord. vals</th>
<th>#X-var</th>
<th>#N-var</th>
<th>#F-var</th>
<th>#S-var</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Wei-Yin Loh 300 GUIDE manual
17.2 Contents of prop30.out

5735 0 5157 9 0 0 35
#P-var #M-var #B-var #C-var #I-var
0 0 0 18 0

Number of cases used for training: 5735
Number of split variables: 53
Number of cases excluded due to 0 W or missing D variable: 0

Constant fitted to cases with missing values in regressor variables
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE's for pruned tree: 0.2500

Simple node models
Equal priors
Unit misclassification costs
Univariate split highest priority
Interaction splits 2nd priority; no linear splits
Split values for N and S variables based on exhaustive search
Maximum number of split levels: 15
Minimum node sample size: 6
Top-ranked variables and 1-df chi-squared values at root node
1 0.3346E+03 cat1
2 0.2728E+03 aps1
3 0.2430E+03 crea1
4 0.2402E+03 meanbp1
: 52 0.1052E+01 meta
53 0.6357E+00 race

Size and CV mean cost and SE of subtrees:

<table>
<thead>
<tr>
<th>Tree</th>
<th>#Nodes</th>
<th>Mean Cost</th>
<th>SE(Mean)</th>
<th>BSE(Mean)</th>
<th>Median Cost</th>
<th>BSE(Median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>354</td>
<td>3.581E-01</td>
<td>6.817E-03</td>
<td>6.202E-03</td>
<td>3.609E-01</td>
<td>5.354E-03</td>
</tr>
<tr>
<td>2</td>
<td>353</td>
<td>3.581E-01</td>
<td>6.817E-03</td>
<td>6.202E-03</td>
<td>3.609E-01</td>
<td>5.354E-03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>208</td>
<td>18</td>
<td>3.278E-01</td>
<td>6.421E-03</td>
<td>4.257E-03</td>
<td>3.277E-01</td>
<td>6.447E-03</td>
</tr>
<tr>
<td>209**</td>
<td>16</td>
<td>3.255E-01</td>
<td>6.349E-03</td>
<td>5.516E-03</td>
<td>3.205E-01</td>
<td>9.186E-03</td>
</tr>
<tr>
<td>210</td>
<td>14</td>
<td>3.287E-01</td>
<td>6.301E-03</td>
<td>5.926E-03</td>
<td>3.290E-01</td>
<td>9.957E-03</td>
</tr>
<tr>
<td>211</td>
<td>12</td>
<td>3.285E-01</td>
<td>6.339E-03</td>
<td>5.849E-03</td>
<td>3.268E-01</td>
<td>8.241E-03</td>
</tr>
<tr>
<td>214</td>
<td>5</td>
<td>3.527E-01</td>
<td>6.506E-03</td>
<td>7.212E-03</td>
<td>3.511E-01</td>
<td>5.489E-03</td>
</tr>
<tr>
<td>215</td>
<td>4</td>
<td>3.690E-01</td>
<td>6.337E-03</td>
<td>7.280E-03</td>
<td>3.706E-01</td>
<td>9.859E-03</td>
</tr>
<tr>
<td>216</td>
<td>2</td>
<td>4.131E-01</td>
<td>5.710E-03</td>
<td>3.745E-03</td>
<td>4.112E-01</td>
<td>3.751E-03</td>
</tr>
<tr>
<td>217</td>
<td>1</td>
<td>5.000E-01</td>
<td>8.419E-03</td>
<td>2.585E-16</td>
<td>5.000E-01</td>
<td>2.764E-16</td>
</tr>
</tbody>
</table>

0-SE tree based on mean is marked with * and has 16 terminal nodes

Wei-Yin Loh

GUIDE manual
17.2  Contents of prop30.out

0-SE tree based on median is marked with + and has 16 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
* tree, ** tree, + tree, and ++ tree all the same

Following tree is based on mean CV with naive SE estimate (**)  

Structure of final tree. Each terminal node is marked with a T.

<table>
<thead>
<tr>
<th>Node</th>
<th>Total cases</th>
<th>Train cases</th>
<th>Predicted class</th>
<th>Node cost</th>
<th>Node split</th>
<th>Interacting variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5735</td>
<td>5735</td>
<td>NoRHC</td>
<td>5.000E-01</td>
<td>cat1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4572</td>
<td>4572</td>
<td>RHC</td>
<td>4.469E-01</td>
<td>pafi1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2218</td>
<td>2218</td>
<td>RHC</td>
<td>3.640E-01</td>
<td>creal1</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>823</td>
<td>823</td>
<td>RHC</td>
<td>4.738E-01</td>
<td>pafi1</td>
<td></td>
</tr>
<tr>
<td>16T</td>
<td>370</td>
<td>370</td>
<td>RHC</td>
<td>3.757E-01</td>
<td>resp</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>453</td>
<td>453</td>
<td>NoRHC</td>
<td>4.385E-01</td>
<td>trauma</td>
<td></td>
</tr>
<tr>
<td>34T</td>
<td>14</td>
<td>14</td>
<td>RHC</td>
<td>9.298E-02</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>439</td>
<td>439</td>
<td>NoRHC</td>
<td>4.193E-01</td>
<td>card</td>
<td></td>
</tr>
<tr>
<td>70T</td>
<td>107</td>
<td>107</td>
<td>RHC</td>
<td>4.213E-01</td>
<td>creal1</td>
<td></td>
</tr>
<tr>
<td>71T</td>
<td>332</td>
<td>332</td>
<td>NoRHC</td>
<td>3.624E-01</td>
<td>bili1:aps1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1395</td>
<td>1395</td>
<td>RHC</td>
<td>3.044E-01</td>
<td>adld3p</td>
<td></td>
</tr>
<tr>
<td>18T</td>
<td>1144</td>
<td>1144</td>
<td>RHC</td>
<td>2.608E-01</td>
<td>wtkilo1</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>251</td>
<td>251</td>
<td>NoRHC</td>
<td>4.675E-01</td>
<td>resp1</td>
<td></td>
</tr>
<tr>
<td>38T</td>
<td>114</td>
<td>114</td>
<td>RHC</td>
<td>3.483E-01</td>
<td>resp1</td>
<td></td>
</tr>
<tr>
<td>39T</td>
<td>137</td>
<td>137</td>
<td>NoRHC</td>
<td>2.852E-01</td>
<td>gastr</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2354</td>
<td>2354</td>
<td>NoRHC</td>
<td>4.682E-01</td>
<td>cat1</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1076</td>
<td>1076</td>
<td>RHC</td>
<td>4.030E-01</td>
<td>meanbp1</td>
<td></td>
</tr>
<tr>
<td>20T</td>
<td>798</td>
<td>798</td>
<td>RHC</td>
<td>3.358E-01</td>
<td>bili1</td>
<td></td>
</tr>
<tr>
<td>21T</td>
<td>278</td>
<td>278</td>
<td>NoRHC</td>
<td>3.753E-01</td>
<td>cat1:age</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1278</td>
<td>1278</td>
<td>NoRHC</td>
<td>3.462E-01</td>
<td>cat2</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>291</td>
<td>291</td>
<td>RHC</td>
<td>4.813E-01</td>
<td>wtkilo1</td>
<td></td>
</tr>
<tr>
<td>44T</td>
<td>108</td>
<td>108</td>
<td>NoRHC</td>
<td>3.287E-01</td>
<td>pafi1</td>
<td></td>
</tr>
<tr>
<td>45T</td>
<td>183</td>
<td>183</td>
<td>RHC</td>
<td>3.834E-01</td>
<td>resp</td>
<td></td>
</tr>
<tr>
<td>23T</td>
<td>987</td>
<td>987</td>
<td>NoRHC</td>
<td>2.898E-01</td>
<td>wtkilo1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1163</td>
<td>1163</td>
<td>NoRHC</td>
<td>2.615E-01</td>
<td>aps1</td>
<td></td>
</tr>
<tr>
<td>6T</td>
<td>895</td>
<td>895</td>
<td>NoRHC</td>
<td>1.666E-01</td>
<td>card</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>268</td>
<td>268</td>
<td>RHC</td>
<td>4.691E-01</td>
<td>cat2</td>
<td></td>
</tr>
<tr>
<td>14T</td>
<td>72</td>
<td>72</td>
<td>RHC</td>
<td>3.052E-01</td>
<td>meanbp1</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>196</td>
<td>196</td>
<td>NoRHC</td>
<td>4.635E-01</td>
<td>income</td>
<td></td>
</tr>
<tr>
<td>30T</td>
<td>25</td>
<td>25</td>
<td>RHC</td>
<td>2.570E-01</td>
<td>wblc1</td>
<td></td>
</tr>
<tr>
<td>31T</td>
<td>171</td>
<td>171</td>
<td>NoRHC</td>
<td>4.154E-01</td>
<td>card</td>
<td></td>
</tr>
</tbody>
</table>

Number of terminal nodes of final tree: 16
Total number of nodes of final tree: 31
Second best split variable (based on curvature test) at root node is \(\text{aps1}\)

Classification tree:
For categorical variable splits, values not in training data go to the right

Node 1: \(\text{cat1} = \text{"ARF"}, \text{"CHF"}, \text{"MOSF w/Malignancy"}, \text{"MOSF w/Sepsis"}\)
    Node 2: \(\text{pafi1} \leq 188.43750\)
        Node 4: \(\text{crea1} \leq 1.2498779\)
            Node 16: RHC
        Node 8: \(\text{pafi1} \leq 116.48438\)
            Node 10: \(\text{meanbp1} \leq 98.500000\)
                Node 21: NoRHC
            Node 20: RHC
            Node 21: NoRHC
        Node 8: \(\text{pafi1} > 116.48438\) or NA
            Node 17: \(\text{trauma} = \text{"Yes"}\)
                Node 34: RHC
            Node 17: \(\text{trauma} /= \text{"Yes"}\)
                Node 35: \(\text{card} = \text{"Yes"}\)
                    Node 70: RHC
                Node 35: \(\text{card} /= \text{"Yes"}\)
                    Node 71: NoRHC
        Node 4: \(\text{crea1} > 1.2498779\) or NA
            Node 9: \(\text{adld3p} = \text{NA}\)
                Node 18: RHC
            Node 9: \(\text{adld3p} /= \text{NA}\)
                Node 19: \(\text{resp1} \leq 29.500000\) or NA
                    Node 38: RHC
                Node 19: \(\text{resp1} > 29.500000\)
                    Node 39: NoRHC
        Node 2: \(\text{pafi1} > 188.43750\) or NA
            Node 5: \(\text{cat1} = \text{"CHF"}, \text{"MOSF w/Sepsis"}\)
                Node 10: \(\text{meanbp1} \leq 98.500000\)
                    Node 21: NoRHC
                Node 20: RHC
                Node 21: NoRHC
            Node 5: \(\text{cat1} /= \text{"CHF"}, \text{"MOSF w/Sepsis"}\)
                Node 11: \(\text{cat2} = \text{"MOSF w/Sepsis"}\)
                    Node 22: \(\text{wtkilo1} \leq 66.449950\)
                        Node 44: NoRHC
                    Node 22: \(\text{wtkilo1} > 66.449950\) or NA
                        Node 45: RHC
                    Node 22: \(\text{wtkilo1} /= \text{NA}\)
                    Node 45: RHC
                Node 11: \(\text{cat2} /= \text{"MOSF w/Sepsis"}\)
                    Node 23: NoRHC
            Node 11: \(\text{cat2} = \text{"MOSF w/Sepsis"}\)
                Node 22: \(\text{wtkilo1} \leq 66.449950\)
                    Node 44: NoRHC
                Node 22: \(\text{wtkilo1} > 66.449950\) or NA
                    Node 45: RHC
                Node 22: \(\text{wtkilo1} /= \text{NA}\)
                Node 45: RHC
            Node 3: \(\text{aps1} \leq 61.500000\)
                Node 6: NoRHC
            Node 3: \(\text{aps1} > 61.500000\) or NA
                Node 7: \(\text{cat2} = \text{"MOSF w/Sepsis"}\)
17.2 Contents of prop30.out

Node 14: RHC
Node 7: cat2 /= "MOSF w/Sepsis"
Node 15: income = ", "$25-50k", "$50k"
Node 30: RHC
Node 15: income /= "$25-50k", "$50k"
Node 31: NoRHC

***************************************************************
Predictor means below are means of cases with no missing values.
Regression coefficients are computed from the complete cases.

Node 1: Intermediate node
A case goes into Node 2 if cat1 = "ARF", "CHF", "MOSF w/Malignancy", "MOSF w/Sepsis"
   cat1 mode = "ARF"
   Number of observations in node = 5735
   Regressor Coefficient t-stat p-value
   Constant 0.3064 38.80 0.000
   swang1.RHC 0.7364E-01 5.756 0.9026E-08
   Number of observations in node = 5735
-----------------------------

Node 2: Intermediate node
A case goes into Node 4 if pafi1 <= 188.43750
   pafi1 mean = 215.63083
   Number of observations in node = 4572
-----------------------------

Node 4: Intermediate node
A case goes into Node 8 if crea1 <= 1.2498779
   crea1 mean = 2.1359302
   Number of observations in node = 2218
-----------------------------

Node 8: Intermediate node
A case goes into Node 16 if pafi1 <= 116.48438
   pafi1 mean = 120.46293
   Number of observations in node = 823
-----------------------------

Node 16: Terminal node
   Regressor Coefficient t-stat p-value
   Constant 0.3115 8.801 0.7772E-15
   swang1.RHC 0.9494E-01 1.907 0.5729E-01
   Number of observations in node = 370
-----------------------------

Node 17: Intermediate node
A case goes into Node 34 if trauma = "Yes"
   trauma mode = "No"
   Number of observations in node = 453
-----------------------------
17.2 Contents of prop30.out : 

----------------------------
Node 30: Terminal node
Regressor Coefficient t-stat p-value 
Constant 0.4444 2.617 0.1543E-01 
swang1.RHC -0.6944E-01 -0.3271 0.7466 
Number of observations in node = 25 
----------------------------

Node 31: Terminal node
Regressor Coefficient t-stat p-value 
Constant 0.5294 11.96 0.000 
swang1.RHC 0.2206 2.748 0.6641E-02 
Number of observations in node = 171 
----------------------------

Regression estimates are weighted means over terminal nodes
Regressor Coefficient z-stat p-value 
constant 0.3160 38.52 0.000 
swang1.RHC 0.5191E-01 3.597 0.3222E-03 

average treatment effect of swang1 level "RHC" vs level "NoRHC" = 5.1909E-02 

Observed and fitted values are stored in prop30.fit
LaTeX code for tree is in prop30.tex
R code is stored in prop30.r

The results at the end of prop30.out show that the average treatment effect is 0.51909. The LaTeX tree is shown in Figure 37. The number beside each terminal node is the proportion of observations with swang1 = RHC (Z = 1). The pair below each node are the sample means of Y corresponding to Z = 0 and 1. GUIDE treats "NoRHC" as Z = 0 because it precedes "RHC" in alphabetical order.

The file prop30.fit gives the proportions of swang1 in the rightmost two columns. Here are the top 5 rows of the file:

<table>
<thead>
<tr>
<th>train</th>
<th>node</th>
<th>observed</th>
<th>predicted</th>
<th>&quot;P(NoRHC)&quot;</th>
<th>&quot;P(RHC)&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>6</td>
<td>&quot;NoRHC&quot;</td>
<td>&quot;NoRHC&quot;</td>
<td>0.89050E+00</td>
<td>0.10950E+00</td>
</tr>
<tr>
<td>y</td>
<td>20</td>
<td>&quot;RHC&quot;</td>
<td>&quot;RHC&quot;</td>
<td>0.45113E+00</td>
<td>0.54887E+00</td>
</tr>
<tr>
<td>y</td>
<td>45</td>
<td>&quot;RHC&quot;</td>
<td>&quot;RHC&quot;</td>
<td>0.50273E+00</td>
<td>0.49727E+00</td>
</tr>
<tr>
<td>y</td>
<td>18</td>
<td>&quot;NoRHC&quot;</td>
<td>&quot;RHC&quot;</td>
<td>0.36451E+00</td>
<td>0.63549E+00</td>
</tr>
<tr>
<td>y</td>
<td>20</td>
<td>&quot;RHC&quot;</td>
<td>&quot;RHC&quot;</td>
<td>0.45113E+00</td>
<td>0.54887E+00</td>
</tr>
</tbody>
</table>

Wei-Yin Loh 305 GUIDE manual
Figure 37: GUIDE v.41.1 0.250-SE tree for propensity score grouping and estimation of effects of \textit{swang1} on \textit{dth30}. At each split, an observation goes to the left branch if and only if the condition is satisfied. Symbol ‘$\leq$’ stands for ‘$\leq$ or missing’. 

\begin{itemize}
  \item $S_1 = \{\text{ARF, CHF, MOSF w/Malignancy, MOSF w/Sepsis}\}$. 
  \item $S_2 = \{\text{CHF, MOSF w/Sepsis}\}$. 
  \item $S_3 = \{\text{MOSF w/Sepsis}\}$. 
  \item $S_4 = \{\text{MOSF w/Sepsis}\}$. 
  \item $S_5 = \{\text{\$25-$50k, > \$50k}\}$. 
\end{itemize}

Circles with dashed lines are nodes with no significant split variables. Sample means of \textit{dth30} for \textit{swang1} levels NoRHC and RHC, respectively, printed below nodes. Sample proportion of \textit{swang1} = RHC printed beside nodes. Second best split variable at root node is \textit{aps1}. 
18 Differential item functioning: GDS data

GUIDE has an experimental option to identify important predictor variables and items with differential item functioning (DIF) in a data set with two or more item (dependent variable) scores. We illustrate it with a data set from Broekman et al. (2011, 2008) and Marc et al. (2008). It consists of responses from 1978 subjects on 15 items. There are 3 predictor variables (age, education, and gender). The data and description files are GDS.dat and GDS.dsc. Although the item responses in this example are 0-1, GUIDE allows them to be in any ordinal (e.g., Likert) scale. The contents of GDS.dsc are:

```
GDS.dat
NA
1
1 rid x
2 satis d
3 drop d
4 empty d
5 bored d
6 spirit d
7 afraid d
8 happy d
9 help d
10 home d
11 memory d
12 alive d
13 worth d
14 energy d
15 hope d
16 better d
17 total x
18 gender c
19 education n
20 age n
21 dxcurrent x
22 sumscore x
```

Here is the session log to create an input file for identifying DIF items and the important predictor variables:

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: dif.in
   Input 1 for model fitting, 2 for importance or DIF scoring, 3 for data conversion ([1:3], <cr>=1): 2

Wei-Yin Loh 307 GUIDE manual
Name of batch output file: dif.out
Input 1 for classification, 2 for regression, 3 for propensity score grouping
Input your choice ([1:3], <cr>=1): 2
Choose type of regression model:
1=linear, 2=quantile, 3=Poisson, 4=censored response,
5=multiresponse or itemresponse, 6=longitudinal data (with T variables).
Input choice ([1:6], <cr>=1): 5
Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
Input name of data description file (max 100 characters);
enclose with matching quotes if it has spaces: GDS.dsc
Reading data description file ...
Training sample file: GDS.dat
Missing value code: NA
Records in data file start on line 1
2 N variables changed to S
Number of D variables: 15
D variables are:
satis
drop
empty
bored
spirit
afraid
happy
help
home
memory
alive
worth
energy
hope
better
Multivariate or univariate split variable selection:
Choose multivariate if there is an order among the D variables;
choose univariate otherwise or if item response
Input 1 for multivariate, 2 for univariate ([1:2], <cr>=1): 2
D variables can be normalized to have unit variance,
e.g., if they have different scales or units
Input 1 to normalize D variables, 2 for no normalization ([1:2], <cr>=1): 2
Input 1 for equal, 2 for unequal weighting of D variables ([1:2], <cr>=1):
Reading data file ... 
Number of records in data file: 1978
Length of longest entry in data file: 4
Checking for missing values ...
Finished checking
Missing values found in D variables
Assigning integer codes to values of 1 categorical variables
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...
Data checks complete
Creating missing value indicators ...
Some D variables have missing values
Rereading data ...
PCA can be used for variable selection
Do not use PCA if differential item functioning (DIF) scores are wanted
Input 1 to use PCA, 2 otherwise ([1:2], <cr>=2):

#cases w/ miss. D = number of cases with all D values missing

<table>
<thead>
<tr>
<th>Total</th>
<th>#cases w/ missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>#cases</td>
<td>miss. D</td>
</tr>
<tr>
<td>1978</td>
<td>0</td>
</tr>
</tbody>
</table>

Number of cases used for training: 1977
Number of split variables: 3
Number of cases excluded due to 0 W or missing D variable: 1
Finished reading data file
Input 1 to save p-value matrix for differential item functioning (DIF), 2 otherwise ([1:2], <cr>=1):
Input file name to store DIF p-values: dif.pv
Input 0=skip LaTeX tree, 1=tree without node numbers, 2=with node numbers ([0:2], <cr>=2):
Input file name to store LaTeX code (use .tex as suffix): dif.tex
You can store the variables and/or values used to split and fit in a file
Choose 1 to skip this step, 2 to store split and fit variables, 3 to store split variables and their values
Input your choice ([1:3], <cr>=1):
You can create a description file with the selected variables included or excluded
Input 2 to create such a file, 1 otherwise ([1:2], <cr>=1):
You can also output the importance scores and variable names to a file
Input 1 to create such a file, 2 otherwise ([1:2], <cr>=1):
Input file name: dif.scr
Input file is created!
Run GUIDE with the command: guide < dif.in

The importance scores are in the file dif.scr. They show that age is most important, followed by gender and education.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Score</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>8.94327E+00</td>
<td>age</td>
</tr>
<tr>
<td>2.00</td>
<td>5.06849E+00</td>
<td>gender</td>
</tr>
<tr>
<td>3.00</td>
<td>3.38749E+00</td>
<td>education</td>
</tr>
</tbody>
</table>

Wei-Yin Loh 309 GUIDE manual
The word 'yes' in the last column of dif.py below shows which item has DIF. In this example, only item #10 (memory) has DIF.

<table>
<thead>
<tr>
<th>Item</th>
<th>Itemname</th>
<th>education</th>
<th>age</th>
<th>gender</th>
<th>DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>satis</td>
<td>0.492E-01</td>
<td>0.399E-01</td>
<td>0.101E+00</td>
<td>no</td>
</tr>
<tr>
<td>2</td>
<td>drop</td>
<td>0.146E-01</td>
<td>0.228E+00</td>
<td>0.923E+00</td>
<td>no</td>
</tr>
<tr>
<td>3</td>
<td>empty</td>
<td>0.207E-02</td>
<td>0.141E+00</td>
<td>0.185E+00</td>
<td>no</td>
</tr>
<tr>
<td>4</td>
<td>bored</td>
<td>0.312E-05</td>
<td>0.212E+00</td>
<td>0.299E+00</td>
<td>no</td>
</tr>
<tr>
<td>5</td>
<td>spirit</td>
<td>0.960E+00</td>
<td>0.737E+00</td>
<td>0.388E-01</td>
<td>no</td>
</tr>
<tr>
<td>6</td>
<td>afraid</td>
<td>0.318E-01</td>
<td>0.472E-03</td>
<td>0.273E-02</td>
<td>no</td>
</tr>
<tr>
<td>7</td>
<td>happy</td>
<td>0.763E+00</td>
<td>0.345E+00</td>
<td>0.251E-01</td>
<td>no</td>
</tr>
<tr>
<td>8</td>
<td>help</td>
<td>0.463E-01</td>
<td>0.611E+00</td>
<td>0.443E-02</td>
<td>no</td>
</tr>
<tr>
<td>9</td>
<td>home</td>
<td>0.371E+00</td>
<td>0.120E+00</td>
<td>0.814E-03</td>
<td>no</td>
</tr>
<tr>
<td>10</td>
<td>memory</td>
<td>0.373E-01</td>
<td>0.000E+00</td>
<td>0.206E-01</td>
<td>yes</td>
</tr>
<tr>
<td>11</td>
<td>alive</td>
<td>0.169E+00</td>
<td>0.155E+00</td>
<td>0.438E+00</td>
<td>no</td>
</tr>
<tr>
<td>12</td>
<td>worth</td>
<td>0.332E+00</td>
<td>0.726E+00</td>
<td>0.696E+00</td>
<td>no</td>
</tr>
<tr>
<td>13</td>
<td>energy</td>
<td>0.660E+00</td>
<td>0.652E+00</td>
<td>0.126E-03</td>
<td>no</td>
</tr>
<tr>
<td>14</td>
<td>hope</td>
<td>0.638E+00</td>
<td>0.392E+00</td>
<td>0.213E+00</td>
<td>no</td>
</tr>
<tr>
<td>15</td>
<td>better</td>
<td>0.517E+00</td>
<td>0.621E+00</td>
<td>0.447E+00</td>
<td>no</td>
</tr>
</tbody>
</table>

Figure 38 shows the tree.
Figure 38: GUIDE v.41.1 importance scoring or DIF regression tree for predicting response variables satis, drop, empty, bored, spirit, afraid, happy, help, home, memory, alive, worth, energy, hope, and better, without using PCA at each node. At each split, an observation goes to the left branch if and only if the condition is satisfied. Sample size (in italics) and predicted values of satis, drop, empty, bored, spirit, afraid, happy, help, home, memory, alive, worth, energy, hope, and better printed below nodes. Second best split variable at root node is gender.
19 Bootstrap confidence intervals

Owing to the numerous procedures that are performed during tree construction (such as selection of the variable and the split set to partition each intermediate node), proper statistical inference must account for the multiple testing and estimation issues. Otherwise, the error variance will be underestimated. Suppose, for example, we wish to obtain confidence intervals for the proportion of “RHC” in each terminal node of the tree in Figure 1. Let \( n \) denote the sample size in a node and \( \hat{p} \) the proportion of observations in it with the response value RHC. The usual \((1 - \alpha)\) binomial interval is then \( \hat{p} \pm z_{1-\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n} \), where \( z_{\alpha} \) is the \( \alpha \)-quantile of the standard normal distribution. This formula yields intervals that are too short because it does not account for the extra variance due to model construction. Bonferroni corrections, which are traditionally used for multiple testing, are inapplicable here because the number of tests are not specified in advance. For example, the number of chi-squared tests at each node depends on the number of variables eligible to split the node and the number of levels of splits depends on the total sample size, extent of pruning, and other parameters such as the minimum sample size in each node.

As with the Bonferroni correction, a natural solution is to change the multiplier \( z_{1-\alpha/2} \) to a larger value. The bootstrap method provides one simple solution. Called “bootstrap calibration”, the procedure is described and analyzed in Loh (1987, 1991) in the context of estimating a nonparametric mean; it is extended to subgroup analysis from regression tree models in Loh et al. (2016, 2019c) and Loh and Zhou (2020). The R code below implements the procedure. It can be used by following these steps:

1. Change the name of the data file (rhcdata.txt here) to realdata.txt.
2. Change the name of the description file (rhcdsc1.txt here) to real.dsc.
3. Change the name of the GUIDE input file (classin.txt here) to real.in.
4. Change the word “RHC” in line 1 of the R code to the name of the desired class in the data file.
5. In Windows, change the word “system” in lines 32, 32, 74 and 75 to “shell” if necessary.
6. Source the program in R.
class.name <- "RHC"  ## name of desired class in realdata.txt
nboot <- 1000
probs <- c(0.80,0.90,0.95,0.98)
zstat <- rep(0,nboot)
### write bootstrap description file boot.dsc
file <- readLines("real.dsc")  ## read real description file
write("bootdata.txt",file="boot.dsc")
len <- length(file)
write(file[2:length(file)],"boot.dsc",append=TRUE)
write(paste(len-2,"w␣w"),"boot.dsc",append=TRUE)
### write bootstrap input file boot.in
file2 <- gsub("real.","boot.",file)  ## replace "real." with "boot."
write(file2,"boot.in")
### read real data
z0 <- read.table("realdata.txt",header=TRUE)
nobs <- nrow(z0)
zt <- cbind(z0,rep(0,nobs))  ### add column of weight 0
write("Bootstrap␣simultaneous␣intervals␣by␣linear␣interpolation␣of␣z",
"results.txt")
write("trials␣␣z80␣␣␣␣z90␣␣␣␣z95␣␣␣␣␣z98␣␣␣␣␣␣bias␣err␣␣␣␣␣sd.err",
"results.txt", append=TRUE)
err.test <- rep(0,nboot)  ## misclassification rates
bias <- 0
for(i in 1:nboot){
    zb <- z0[sample(nobs,nobs,replace=TRUE),]
    zb <- cbind(zb,rep(1,nobs))  ### add column of weight 1
write.table(zb,"bootdata.txt",col.names=TRUE,row.names=FALSE)
write.table(zt,"bootdata.txt",col.names=FALSE,row.names=FALSE,
append=TRUE)
system("rm␣-f␣log.txt␣boot.out␣boot.fit")
system("guide␣<␣boot.in␣>␣log.txt")
bfit <- read.table("boot.fit",header=TRUE)  ## read boot results
bias <- bias+(err.resub-err.test[i])
unodes <- unique(sort(bfit$node))
for(j in 1:length(unodes)){
gp <- bfit$node == unodes[j] & bfit$train == "y"  ## training data
n0 <- sum(bfit$observed[gp] != class.name)
n1 <- sum(bfit$observed[gp] == class.name)
ntot <- n0+n1
estp <- n1/ntot
if(n1 == 0 | n0 == 0){
p <- (n1+0.5)/(ntot+1)
sd <- sqrt(p*(1-p)/(ntot+1))
} else {
  sd <- sqrt(estp*(1-estp)/ntot)
}

gp <- bfit$node == unodes[j] & bfit$train == "n"  ## real data
n0 <- sum(bfit$observed[gp] != class.name)
np <- sum(bfit$observed[gp] == class.name)
realp <- np/(n0+np)

zstat[i] <- max(zstat[i],abs(realp-estp)/sd)

if(i %% 100 == 0){
  sd.err <- sqrt(var(err.test[1:i]))  ## linear interpolation
  q <- quantile(zstat[1:i],probs=probs,type=4)
  write(c(i,q,bias/sd.err),"results.txt",append=TRUE,ncol=7)
}

### find calibrated z. alpha
write(paste("No.␣bootstraps␣=␣",nboot),"results.txt",append=TRUE)
write(c("Calibrated␣z␣at␣levels",probs),file="results.txt",ncol=5,
append=TRUE)
q <- quantile(zstat,probs=probs,type=4)  ## linear interpolation
write(q,"results.txt",append=TRUE,ncol=4)
write(paste("Bootstrap␣estimate␣of␣bias␣of␣error␣rate",bias/nboot),
"results.txt",append=TRUE)
write(paste("Bootstrap␣estimate␣of␣SD␣of␣error␣rate",sqrt(var(err.test))),"results.txt",append=TRUE)

## fit real data
system("rm␣-f␣log.txt␣real.out␣real.fit")
realfit <- read.table("real.fit",header=TRUE)
err.obs <- sum(realfit$observed[train] != realfit$predicted[train])/nobs
write(paste("Real␣data␣observed␣error␣rate␣=␣",err.obs),"results.txt",append=TRUE)

k <- 3  ## 95% level
z0 <- q[k]  ## 95% z value
write(c("Simultaneous␣intervals␣at␣level",probs[k]),
file="results.txt",ncol=2,append=TRUE)
write(paste0("Node␣="unodes," halfwidth left right"),
"results.txt",append=TRUE)
unodes <- unique(sort(realfit$node))
for(j in 1:length(unodes)){
  gp <- realfit$node == unodes[j] & realfit$train == "y"
  n0 <- sum(realfit$observed[gp] != class.name)
n1 <- sum(realfit$observed[gp] == class.name)
  ntot <- n0+n1

  if(i %% 100 == 0){
    sd.err <- sqrt(var(err.test[1:i]))  ## linear interpolation
    q <- quantile(zstat[1:i],probs=probs,type=4)
    write(c(i,q,bias/sd.err),"results.txt",append=TRUE,ncol=7)
  }
}

Wei-Yin Loh 314 GUIDE manual
if(n1 == 0 | n0 == 0){
  p <- (n1+0.5)/(ntot+1)
 .sd <- sqrt(p*(1-p)/(ntot+1))
} else {
  p <- n1/ntot
  sd <- sqrt(p*(1-p)/(ntot))
}

p <- n1/ntot
halfwid <- z0*sd
left <- p-halfwid
rght <- p+halfwid
write(c(unodes[j],ntot,p,halfwid,left,rght),"results.txt",
append=TRUE,ncol=6)

## write(sort(zstat),"zstat.txt",ncol=1) ## output sorted zstat values

Figure 39 gives the contents of the file results.txt. It shows that the calibrated z-multiplier is 3.961722, 4.325215, 4.690964, or 5.337637 for 80%, 90%, 95%, or 98% simultaneous confidence intervals. For 95% intervals, the left and right end points of the intervals in each terminal node are given in the bottom half of the file. These intervals are printed below the terminal nodes in Figure 40.

20 Tree ensembles

A tree ensemble is a collection of trees. GUIDE has two methods of constructing an ensemble.

GUIDE forest. This the preferred method. Similar to Random Forest (Breiman, 2001), it fits unpruned trees to bootstrap samples and randomly selects a small subset of variables to search for splits at each node. There are, however, two important differences:

1. GUIDE forest uses the unbiased GUIDE method for split selection; Random Forest uses the biased CART method. One consequence is that GUIDE forest can be very much faster than Random Forest if the dependent variable is a class variable having more than two distinct values and some categorical predictor variables have many categories.

2. GUIDE forest is applicable to data with missing values. The R implementation of Random Forest (Liaw and Wiener, 2002) requires apriori imputation of missing values in the predictor variables.
Bootstrap simultaneous intervals by linear interpolation of z

<table>
<thead>
<tr>
<th>Trials</th>
<th>z80</th>
<th>z90</th>
<th>z95</th>
<th>z98</th>
<th>bias.err</th>
<th>sd.err</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>4.036962</td>
<td>4.458809</td>
<td>4.545827</td>
<td>4.922293</td>
<td>--0.03357803</td>
<td>0.005906056</td>
</tr>
<tr>
<td>200</td>
<td>4.123996</td>
<td>4.508203</td>
<td>4.777955</td>
<td>5.035208</td>
<td>--0.03335222</td>
<td>0.005670584</td>
</tr>
<tr>
<td>300</td>
<td>4.093978</td>
<td>4.513735</td>
<td>4.918732</td>
<td>5.117146</td>
<td>--0.0335048</td>
<td>0.00598086</td>
</tr>
<tr>
<td>400</td>
<td>4.108083</td>
<td>4.519645</td>
<td>4.835633</td>
<td>5.28808</td>
<td>--0.03360811</td>
<td>0.005930667</td>
</tr>
<tr>
<td>500</td>
<td>4.108083</td>
<td>4.508203</td>
<td>4.826329</td>
<td>5.117146</td>
<td>--0.03377507</td>
<td>0.005887693</td>
</tr>
<tr>
<td>600</td>
<td>4.14132</td>
<td>4.548011</td>
<td>4.895352</td>
<td>5.408027</td>
<td>--0.03397879</td>
<td>0.005812075</td>
</tr>
<tr>
<td>700</td>
<td>4.12996</td>
<td>4.529434</td>
<td>4.889087</td>
<td>5.408027</td>
<td>--0.03377357</td>
<td>0.005839512</td>
</tr>
<tr>
<td>800</td>
<td>4.117319</td>
<td>4.51814</td>
<td>4.845685</td>
<td>5.365021</td>
<td>--0.03369159</td>
<td>0.00588305</td>
</tr>
<tr>
<td>900</td>
<td>4.108552</td>
<td>4.50332</td>
<td>4.835633</td>
<td>5.408027</td>
<td>--0.03358888</td>
<td>0.005924705</td>
</tr>
<tr>
<td>1000</td>
<td>4.108083</td>
<td>4.495735</td>
<td>4.845685</td>
<td>5.397256</td>
<td>--0.03353304</td>
<td>0.005951228</td>
</tr>
</tbody>
</table>

No. bootstraps = 1000
Calibrated z at levels 0.8 0.9 0.95 0.98
4.108083 4.495735 4.845685 5.397256

Bootstrap estimate of bias of error rate = -0.0335330427201395
Bootstrap estimate of SD of error rate = 0.00595122775778847
Real data observed error rate = 0.296251089799477

Simultaneous intervals at level 0.95

<table>
<thead>
<tr>
<th>Node N</th>
<th>P(RHC) halfwid</th>
<th>left</th>
<th>right</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.3816254</td>
<td>0.0989446</td>
<td>0.282681</td>
</tr>
<tr>
<td>7</td>
<td>0.2355072</td>
<td>0.03913718</td>
<td>0.1963701</td>
</tr>
<tr>
<td>8</td>
<td>0.6961832</td>
<td>0.08707675</td>
<td>0.6091065</td>
</tr>
<tr>
<td>18</td>
<td>0.6270492</td>
<td>0.1500158</td>
<td>0.4770334</td>
</tr>
<tr>
<td>19</td>
<td>0.3853211</td>
<td>0.1597212</td>
<td>0.2255999</td>
</tr>
<tr>
<td>25</td>
<td>0.3484848</td>
<td>0.2842088</td>
<td>0.06427609</td>
</tr>
<tr>
<td>26</td>
<td>0.6363636</td>
<td>0.2222518</td>
<td>0.4141119</td>
</tr>
<tr>
<td>27</td>
<td>0.3627888</td>
<td>0.09503228</td>
<td>0.2676965</td>
</tr>
<tr>
<td>48</td>
<td>0.6552511</td>
<td>0.1100458</td>
<td>0.5452053</td>
</tr>
<tr>
<td>49</td>
<td>0.3506494</td>
<td>0.2635033</td>
<td>0.08714608</td>
</tr>
</tbody>
</table>

Figure 39: Contents of results.txt
Figure 40: GUIDE v.38.0 0.25-SE classification tree for predicting swang1 using estimated priors and unit misclassification costs. At each split, an observation goes to the left branch if and only if the condition is satisfied. Symbol ‘$\leq^*$’ stands for ‘$\leq$’ or missing’. Set $S_1 = \{\text{CHF, MOSF w/Sepsis}\}$. Set $S_2 = \{\text{No insurance, Private, Private & Medicare}\}$. Set $S_3 = \{\text{ARF, Lung Cancer, MOSF w/Malignancy}\}$. Predicted classes and sample sizes printed below terminal nodes; class sample proportion for swang1 = RHC beside nodes. Bootstrap calibrated 95% simultaneous intervals for proportion of RHC below nodes.
The default number of trees for GUIDE forest is 1000 if there are fewer than 500 training samples and 100 predictor variables; otherwise, the default is 500.

**Bagged GUIDE.** This fits pruned GUIDE trees to bootstrap samples of the training data (Breiman, 1996). Each tree is pruned by 5-fold cross-validation. The default number of trees is 200 if there are fewer than 500 training samples and 100 predictor variables; otherwise, the default is 100.

With the default settings, GUIDE forest is typically much faster than bagged GUIDE.

### 20.1 GUIDE forest: CE data

#### 20.1.1 Input file creation

0. Read the warranty disclaimer
1. Create a GUIDE input file
   
   Input your choice: 1
   
   Name of batch input file: gf.in
   
   Input 1 for model fitting, 2 for importance or DIF scoring, 3 for data conversion ([1:3], <cr>=1):
   
   Name of batch output file: gf.out
   
   Input 1 for single tree, 2 for ensemble ([1:2], <cr>=1): 2
   
   Input 1 for bagging, 2 for rforest: ([1:2], <cr>=2):
   
   Input 1 for classification, 2 for least-squares regression
   
   Input your choice ([1:2], <cr>=1): 2
   
   Input 1 for default options, 2 otherwise ([1:2], <cr>=1):
   
   Input name of data description file (max 100 characters); enclose with matching quotes if it has spaces: ce2021class.dsc

Reading data description file ...

Training sample file: ce2021.txt

Missing value code: NA

Records in data file start on line 2

Number of M variables associated with C variables: 19

384 N variables changed to S

D variable is INTRDVX_

Reading data file ...

Number of records in data file: 3965

Length of longest entry in data file: 11

Checking for missing values ...

Finished checking

Missing values found among categorical variables

Separate categories will be created for missing categorical variables

Missing values found among non-categorical variables

Recoding D values to integers

Finished recoding
Number of classes: 3
Finding number of levels of M variables associated with C variables ...
Assigning integer codes to values of 47 categorical variables
Finished assigning codes to 10 categorical variables
Finished assigning codes to 20 categorical variables
Finished assigning codes to 30 categorical variables
Finished assigning codes to 40 categorical variables
Associating missing values of N and S variables with M variable codes ...
Re-checking data ...
Allocating missing value information ...
Assigning codes to missing values, if any ...
Data checks complete
Creating missing value indicators ...
Rereading data ...
Warning: S variable DIRACC is constant
Warning: S variable TOTHVHRP is constant
Warning: S variable TOTHVHRC is constant
Warning: S variable ROTHRFLC is constant
Warning: S variable WELFREBX is constant
Smallest positive weight: 1.0725E+03
Largest positive weight: 9.3902E+04
Class #Cases Proportion
C 1478 0.37276166
D 2431 0.61311475
T 56 0.01412358

Total #cases w/ #missing
#cases miss. D ord. vals #X-var #N-var #F-var #S-var
3965 0 3965 1 0 0 384
#P-var #M-var #B-var #C-var #I-var
0 116 0 47 0
Number of cases used for training: 3965
Number of split variables: 431
Number of cases excluded due to 0 W or missing D variable: 0
Finished reading data file
Warning: No linear splits; number of S variables must be < 225
Choose 1 for estimated priors, 2 for equal priors, 3 to input priors from a file
Input 1, 2, or 3 ([1:3], <cr>=1):
Choose 1 for unit misclassification costs, 2 to input costs from a file
Input 1 or 2 ([1:2], <cr>=1):
Warning: All positive weights treated as 1
Input name of file to store predicted class and probability: gf.pro
Input rank of top variable to split root node ([1:431], <cr>=1):
Input file is created!
Run GUIDE with the command: guide < gf.in
20.1.2 Contents of gf.out

Note: Owing to the intrinsic randomness in forests, your results may differ from those shown below. “OOB” stands for “out-of-bag”.

Random forest of classification trees
No pruning
Data description file: ce2021class.dsc
Training sample file: ce2021.txt
Missing value code: NA
Records in data file start on line 2
Number of M variables associated with C variables: 19
384 N variables changed to S
D variable is INTRDVX_
Number of records in data file: 3965
Length of longest entry in data file: 11
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Number of classes: 3
Warning: S variable DIRACC is constant
Warning: S variable TOTHVHRP is constant
Warning: S variable TOTHVHRC is constant
Warning: S variable ROTHRFLC is constant
Warning: S variable WELFREBX is constant
Smallest and largest positive weights are 1.0725E+03 and 9.3902E+04
Training sample class proportions of D variable INTRDVX_:
Class   #Cases    Proportion
C       1478   0.37276166
D       2431   0.61311475
T        56   0.01412358

Summary information for training sample of size 3965
d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators),
s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
m=missing-value flag variable, p=periodic variable, w=weight
Levels of M variables are for missing values in associated variables

<table>
<thead>
<tr>
<th></th>
<th>Codes/Levels</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Periods</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DIRACC</td>
<td>s</td>
<td>1.000</td>
<td>1.000</td>
<td>170</td>
</tr>
<tr>
<td>2</td>
<td>DIRACC_</td>
<td>m</td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>AGE_REF</td>
<td>s</td>
<td>18.00</td>
<td>87.00</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>AGE_REF_</td>
<td>m</td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>AGE2</td>
<td>s</td>
<td>21.00</td>
<td>87.00</td>
<td>1734</td>
</tr>
<tr>
<td>6</td>
<td>AGE2_</td>
<td>m</td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

Wei-Yin Loh

GUIDE manual
### GUIDE forest: CE data

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Code</th>
<th>Value min</th>
<th>Value max</th>
</tr>
</thead>
<tbody>
<tr>
<td>STOCKYR</td>
<td>c</td>
<td>1</td>
<td>3942</td>
<td></td>
</tr>
<tr>
<td>STOCKYR_</td>
<td>m</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WHLFYR</td>
<td>c</td>
<td>1</td>
<td>3964</td>
<td></td>
</tr>
<tr>
<td>WHLFYR_</td>
<td>m</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FFTAXOWE</td>
<td>s</td>
<td>-0.3368E+05</td>
<td>0.3997E+06</td>
<td></td>
</tr>
<tr>
<td>FSTAXOWE</td>
<td>s</td>
<td>-3309</td>
<td>0.7223E+05</td>
<td></td>
</tr>
</tbody>
</table>

| Total | #cases w/ missing | #cases miss. D ord. vals | #X-var | #N-var | #F-var | #S-var | #P-var | #M-var | #B-var | #C-var | #I-var |
|-------|-------------------|--------------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
|       |                   |                          |        |        |        |        |        |        |        |        |        |        |
|       | 3965              | 0                        | 3965   | 1      | 0      | 0      | 384    | 0      | 116    | 0      | 47     | 0      |

Number of cases used for training: 3965
Number of split variables: 431
Number of cases excluded due to 0 W or missing D variable: 0

Number of trees in ensemble: 500
Number of variables used for splitting: 144
Warning: No linear splits; number of S variables must be < 225
Simple node models
Estimated priors
Unit misclassification costs
Warning: All positive weights treated as 1
Univariate split highest priority
No interaction splits
No linear splits
Fraction of cases used for splitting each node: .0252
Maximum number of split levels: 18
Minimum node sample size: 19
Mean number of terminal nodes: 135.8

#### Classification matrix for training sample:

<table>
<thead>
<tr>
<th>Predicted class</th>
<th>True class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
</tr>
<tr>
<td>C</td>
<td>1047</td>
</tr>
<tr>
<td>D</td>
<td>431</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
</tr>
</tbody>
</table>

Total: 1478 2431 56

Number of cases used for tree construction: 3965
Number misclassified: 541
Resubstitution estimate of mean misclassification cost: .1364

Number of OOB cases: 3965
Number OOB misclassified: 961

Wei-Yin Loh 321 GUIDE manual
0.2424
Mean number of trees per OOB observation: 183.93

Predicted class probabilities are stored in gf.pro

Following are the top few rows of the file gf.pro, which give the estimated class posterior probabilities and the predicted and observed values of each case in the data.

<table>
<thead>
<tr>
<th>train</th>
<th>&quot;P(C)&quot;</th>
<th>&quot;P(D)&quot;</th>
<th>&quot;P(T)&quot;</th>
<th>predicted</th>
<th>observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>0.79713E+00</td>
<td>0.19766E+00</td>
<td>0.52107E-02</td>
<td>&quot;C&quot;</td>
<td>&quot;C&quot;</td>
</tr>
<tr>
<td>y</td>
<td>0.17413E+00</td>
<td>0.81800E+00</td>
<td>0.78689E-02</td>
<td>&quot;D&quot;</td>
<td>&quot;D&quot;</td>
</tr>
<tr>
<td>y</td>
<td>0.19204E+00</td>
<td>0.80242E+00</td>
<td>0.55365E-02</td>
<td>&quot;D&quot;</td>
<td>&quot;D&quot;</td>
</tr>
<tr>
<td>y</td>
<td>0.30491E+00</td>
<td>0.69031E+00</td>
<td>0.47743E-02</td>
<td>&quot;D&quot;</td>
<td>&quot;D&quot;</td>
</tr>
<tr>
<td>y</td>
<td>0.28991E+00</td>
<td>0.69812E+00</td>
<td>0.11970E-01</td>
<td>&quot;D&quot;</td>
<td>&quot;D&quot;</td>
</tr>
<tr>
<td>y</td>
<td>0.40932E+00</td>
<td>0.48080E+00</td>
<td>0.10988E+00</td>
<td>&quot;D&quot;</td>
<td>&quot;T&quot;</td>
</tr>
<tr>
<td>y</td>
<td>0.36835E+00</td>
<td>0.61102E+00</td>
<td>0.20625E-01</td>
<td>&quot;D&quot;</td>
<td>&quot;D&quot;</td>
</tr>
</tbody>
</table>

20.2 Bagged GUIDE

This option uses an ensemble of pruned GUIDE trees. It often takes longer to execute and does not appear to produce more accurate results. It is made available for research purposes.

21 Other features

21.1 Pruning with test samples

GUIDE typically has three pruning options for deciding the size of the final tree: (i) cross-validation, (ii) test sample, and (iii) no pruning. Test-sample pruning is available only when there are no derived variables, such as creation of dummy indicator variables when ‘b’ variables are present. If test-sample pruning is chosen, the program will ask for the name of the file containing the test samples. This file must have the same column format as the training sample file. Pruning with test-samples or no pruning are non-default options.

21.2 Prediction of test samples

GUIDE can produce R code to predict future observations from all except kernel and nearest neighbor classification and ensemble models. This is also a non-default option.
Predictions of the training data for all models can be obtained, however, at the time of tree construction. This feature can be used to obtain predictions on “test samples” (i.e., observations that are not used in tree construction) by adding them to the training sample file. There are two ways to distinguish the test observations from the training observations:

1. Use a weight variable (designated as $W$ in the description file) that takes value 1 for each training observation and 0 or each test observation.

2. Replace the $D$ values of the test observations with the missing value code.

For tree construction, GUIDE does not use observations in the training sample file that have zero weight.

### 21.3 GUIDE in R and in simulations

GUIDE can be used in simulations or used repeatedly on bootstrap samples to produce an ensemble of tree models. For the latter,

1. Create a file (with name `data.txt`, say) containing one set of bootstrapped data.

2. Create a data description file (with name `desc.txt`, say) that refers to `data.txt`.

3. Create an input file (with name `input.txt`, say) that refers to `desc.txt`.

4. Write a batch program (Windows) or a shell script (Linux or Macintosh) that repeatedly:

   (a) replaces the file `data.txt` with new bootstrapped samples;
   (b) calls GUIDE with the command: `guide < input.txt`; and
   (c) reads and processes the results from each GUIDE run.

In R, the command in step 4b depends on the operating system. If the GUIDE program and the files `data.txt` and `input.txt` are in the same folder as the working R directory, the command is:

**Linux/Macintosh:** `system("guide < input.txt > log.txt")`

**Windows:** `shell("guide < input.txt > log.txt")`

If the files are not all in the same folder, full path names must be given. Here `log.txt` is a text file that stores messages during execution. If GUIDE does not run successfully, errors are also written to `log.txt`. 
21.4 Generation of powers and products

GUIDE allows the creation of certain powers and products of regressor variables on the fly. Specifically, variables of the form $X_1^pX_2^q$, where $X_1$ and $X_2$ are numerical predictor variables and $p$ and $q$ are integers, can be created by adding one or more lines of the form

```
0 i p j q a
```

at the end of the data description file. Here $i$ and $j$ are integers giving the column numbers of variables $X_1$ and $X_2$, respectively, in the data file and $a$ is one of the letters $n$, $s$, or $f$ (corresponding to a numerical variable used for both splitting and fitting, splitting only, or fitting only).

To demonstrate, suppose we wish to fit a piecewise quadratic model in the variable `wtgain` in the birthweight data. This is easily done by adding one line to the file `birthwt.dsc`. First we assign the $s$ (for splitting only) designator to every numerical predictor except `wtgain`. This will prevent all variables other than `wtgain` from acting as regressors in the piecewise quadratic models. To create the variable `wtgain^2`, add the line

```
0 8 2 8 0 f
```

to the end of `birthwt.dsc`. The 8’s in the above line refer to the column number of the variables `wtgain` in the data file, and the $f$ tells the program to use the variable `wtgain^2` for fitting terminal node models only. Note: The line defines `wtgain^2` as `wtgain^2 = wtgain^1 * wtgain^0`. Since we can equivalently define the variable by `wtgain^2 = wtgain^1 * wtgain^1`, we could also have used the line: “0 8 1 8 1 f”.

The resulting description file now looks like this:

```
birthwt.dat
NA
1
1 weight d
2 black c
3 married c
4 boy c
5 age s
6 smoke c
7 cigsper s
8 wtgain n
9 visit c
10 ed c
11 lowbwt x
0 8 2 8 0 f
```
When the program is given this description file, the output will show the regression coefficients of $\text{wtgain}$ and $\text{wtgain}^2$ in each terminal node of the tree.

### 21.5 Data formatting functions

GUIDE has a utility function for reformatting data files into forms required by some old statistical software packages:

1. **R/Splus**: Fields are space delimited. Missing values are coded as `NA`. Each record is written on one line. Variable names are given on the first line.

2. **SAS**: Fields are space delimited. Missing values are coded with periods. Character strings are truncated to eight characters. Spaces within character strings are replaced with underscores (`_`).

3. **TEXT**: Fields are comma delimited. Empty fields denote missing values. Character strings longer than eight characters are truncated. Each record is written on one line. Variable names are given on the first line.

4. **STATISTICA**: Fields are comma delimited. Commas in character strings are stripped. Empty fields denote missing values. Each record occupies one line.

5. **SYSTAT**: Fields are comma delimited. Strings are truncated to eight characters. Missing character values are replaced with spaces, missing numerical values with periods. Each record occupies one line.

6. **BMDP**: Fields are space delimited. Categorical values are sorted in alphabetic order and then assigned integer codes. Missing values are indicated by asterisks. Variable names longer than eight characters are truncated.

7. **DataDesk**: Fields are space delimited. Missing categorical values are coded with question marks. Missing numerical values are coded with asterisks. Each record is written on one line. Spaces within categorical values are replaced with underscores. Variable names are given on the first line of the file.

8. **MINITAB**: Fields are space delimited. Categorical values are sorted in alphabetic order and then assigned integer codes. Missing values are coded with asterisks. Variable names longer than eight characters are truncated.

9. **NUMBERS**: Same as **TEXT** option except that categorical values are converted to integer codes.
10. C4.5: This is the format required by the C4.5 (Quinlan, 1993) program.

11. ARFF: This is the format required by the WEKA (Witten and Frank, 2000) programs.

Following is a sample session where the NHTSA comma-separated data are reformatted to tab-delimited for R or Splus.

0. Read the warranty disclaimer
1. Create a GUIDE input file
   Input your choice: 1
   Name of batch input file: format.in
   Input 1 for model fitting, 2 for importance or DIF scoring, 3 for data conversion ([1:3], <cr>=1): 3
   Name of batch output file: format.out
   Input 1 if D variable is categorical, 2 if real ([1:2], <cr>=1):
   Input name of data description file (max 100 characters); enclose with matching quotes if it has spaces: nhtsaclass.dsc
   nhtsaclass.dsc
   Reading data description file ...
   Training sample file: nhtsadata.csv
   Missing value code: NA
   Records in data file start on line 2
   Warning: 48 N variables changed to S
   Dependent variable is HIC2
   Reading data file ...
   Number of records in data file: 3310
   Length of longest entry in data file: 19
   Checking for missing values ...
   Total number of cases: 3310
   Number of classes: 2

   Warning: "x" variables will be excluded
   Choose one of the following data formats:

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
<th>Separ</th>
<th>char.</th>
<th>numer.</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>R/Splus</td>
<td>space</td>
<td>NA</td>
<td>NA</td>
<td>1 line/case, var names on 1st line</td>
</tr>
<tr>
<td>2</td>
<td>SAS</td>
<td>space</td>
<td>.</td>
<td>.</td>
<td>strings trunc., spaces -&gt; ' '</td>
</tr>
<tr>
<td>3</td>
<td>TEXT</td>
<td>comma</td>
<td>empty</td>
<td>empty</td>
<td>1 line/case, var names on 1st line</td>
</tr>
<tr>
<td>4</td>
<td>STATISTICA</td>
<td>comma</td>
<td>empty</td>
<td>empty</td>
<td>1 line/case, commas stripped</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>var names on 1st line</td>
</tr>
<tr>
<td>5</td>
<td>SYSTAT</td>
<td>comma</td>
<td>space</td>
<td>.</td>
<td>1 line/case, var names on 1st line</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>strings trunc. to 8 chars</td>
</tr>
<tr>
<td>6</td>
<td>BMDP</td>
<td>space</td>
<td>*</td>
<td></td>
<td>strings trunc. to 8 chars</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cat values -&gt; integers (alph. order)</td>
</tr>
</tbody>
</table>

Wei-Yin Loh 326 GUIDE manual
21.5 Data formatting functions

<table>
<thead>
<tr>
<th>Format</th>
<th>Space/Comma</th>
<th>NA</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 DATADESK</td>
<td>space</td>
<td>?</td>
<td>1 line/case, var names on 1st line, spaces -&gt; '_'</td>
</tr>
<tr>
<td>8 MINITAB</td>
<td>space</td>
<td>*</td>
<td>cat values -&gt; integers (alph. order), var names trunc. to 8 chars</td>
</tr>
<tr>
<td>9 NUMBERS</td>
<td>comma</td>
<td>NA</td>
<td>1 line/case, var names on 1st line, cat values -&gt; integers (alph. order)</td>
</tr>
<tr>
<td>10 C4.5</td>
<td>comma</td>
<td>?</td>
<td>1 line/case, dependent variable last</td>
</tr>
<tr>
<td>11 ARFF</td>
<td>comma</td>
<td>?</td>
<td>1 line/case</td>
</tr>
</tbody>
</table>

-----------------------------
0 abort this job

Input your choice ([0:11], <cr>=1):
Input name of new data file: newdata.txt
Input file is created!
Run GUIDE with the command: guide < format.in
## A CE variables

Table 11: Some CE variables and their missing rates (if any)

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE_REF</td>
<td>Age of reference person</td>
<td></td>
</tr>
<tr>
<td>AGE2</td>
<td>Age of spouse</td>
<td>0.44</td>
</tr>
<tr>
<td>ALCBEVCQ</td>
<td>Alcoholic beverages this quarter</td>
<td></td>
</tr>
<tr>
<td>ALCBEVPQ</td>
<td>Alcoholic beverages last quarter</td>
<td></td>
</tr>
<tr>
<td>ALLFULCQ</td>
<td>Fuel oil and other fuels this quarter</td>
<td></td>
</tr>
<tr>
<td>ALLFULPQ</td>
<td>Fuel oil and other fuels last quarter</td>
<td></td>
</tr>
<tr>
<td>APPARCQ</td>
<td>Apparel and services this quarter (MENBOYCQ + WOMGRLCQ + CHLDRNCQ + FOOTWRCQ + OTH-APLCQ)</td>
<td></td>
</tr>
<tr>
<td>APPARPQ</td>
<td>Apparel and services last quarter (same composition as APPARCQ)</td>
<td></td>
</tr>
<tr>
<td>AS_COMP1</td>
<td>Number of males age 16 and over in CU</td>
<td></td>
</tr>
<tr>
<td>AS_COMP2</td>
<td>Number of females age 16 and over in CU</td>
<td></td>
</tr>
<tr>
<td>AS_COMP3</td>
<td>Number of males age 2 through 15 in CU</td>
<td></td>
</tr>
<tr>
<td>AS_COMP4</td>
<td>Number of females age 2 through 15 in CU</td>
<td></td>
</tr>
<tr>
<td>AS_COMP5</td>
<td>Number of members under age 2 in CU</td>
<td></td>
</tr>
<tr>
<td>BATHRMQ</td>
<td>Number of complete baths in this unit</td>
<td>0.01</td>
</tr>
<tr>
<td>BBYDAYCQ</td>
<td>Babysitting and child day care this quarter</td>
<td></td>
</tr>
<tr>
<td>BBYDAYPQ</td>
<td>Babysitting and child day care last quarter</td>
<td></td>
</tr>
<tr>
<td>BEDROOMQ</td>
<td>Number of bedrooms in CU</td>
<td></td>
</tr>
<tr>
<td>BLS_URBN</td>
<td>Is this CU located in an urban or rural area? (1=urban, 2=rural)</td>
<td></td>
</tr>
<tr>
<td>BUILDING</td>
<td>Which of these descriptions from the list best describes this building? (1–11)</td>
<td></td>
</tr>
<tr>
<td>BUILT</td>
<td>Year property was built</td>
<td>0.23</td>
</tr>
<tr>
<td>BUSCREEN</td>
<td>Has household had business expenses that could be reimbursed? (1=yes, 2=no)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>CARTKNCQ</td>
<td>Cars and trucks, new (net outlay) this quarter</td>
<td></td>
</tr>
<tr>
<td>CARTKNPQ</td>
<td>Cars and trucks, new (net outlay) last quarter</td>
<td></td>
</tr>
<tr>
<td>CARTKUCQ</td>
<td>Cars and trucks, used (net outlay) this quarter</td>
<td></td>
</tr>
<tr>
<td>CARTKUPQ</td>
<td>Cars and trucks, used (net outlay) last quarter</td>
<td></td>
</tr>
<tr>
<td>CASHCOCQ</td>
<td>Cash contributions this quarter</td>
<td></td>
</tr>
<tr>
<td>CASHCOPQ</td>
<td>Cash contributions last quarter</td>
<td></td>
</tr>
</tbody>
</table>

Wei-Yin Loh 328 GUIDE manual
<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHILDAGE</td>
<td>Age of children of reference person (0=no children, 1=all children less than 6, 2=oldest child 6–11 and at least one child less than 6, 3=all children 6–11, 4=oldest child 12–17 and at least one child less than 12, 5=all children 12–17, 6=oldest child greater than 17 and at least one child less than 17, 7=all children greater than 17)</td>
<td></td>
</tr>
<tr>
<td>CREDFINX</td>
<td>What was the total amount paid in finance, late charges, and interest for all cards in last month?</td>
<td>0.82</td>
</tr>
<tr>
<td>CREDITB</td>
<td>Could you tell me which range that best reflects the total amount owed on all major credit cards including store cards and gas cards? (1=0–499, 2=500–999, 3=1000–2499, 4=2500–9999, 5=10000–34999, 6=35K and over)</td>
<td>0.99</td>
</tr>
<tr>
<td>CREDITBX</td>
<td>Median bracket range of CREDITB</td>
<td>0.99</td>
</tr>
<tr>
<td>CREDITX</td>
<td>Total amount owed on all cards</td>
<td>0.81</td>
</tr>
<tr>
<td>CREDITYRX</td>
<td>Total amount owed on all cards one year ago today</td>
<td>0.90</td>
</tr>
<tr>
<td>CREDYR</td>
<td>Did you have any credit cards including store cards and gas cards one year ago today? (1=yes, 2=no)</td>
<td>0.99</td>
</tr>
<tr>
<td>CREDYRB</td>
<td>Range that best reflects the total amount owed on all major credit cards including store cards and gas cards one year ago today (1=0–499, 2=500–999, 3=1000–2499, 4=2500–9999, 5=10000–34999, 6=35K and over)</td>
<td>0.99</td>
</tr>
<tr>
<td>CREDYRBX</td>
<td>Median bracket range of CREDYRB</td>
<td>0.99</td>
</tr>
<tr>
<td>CUTFNRE</td>
<td>Housing tenure (1=homeowner with mortgage, 2=homeowner without mortgage, 3=homeowner, mortgage not reported, 4=rented, 5=occupied without payment of rent, 6=student housing)</td>
<td></td>
</tr>
<tr>
<td>DEFBENRP</td>
<td>Do you have a defined retirement plan, such as a pension, from an employer? (1=yes, 2=no)</td>
<td>0.77</td>
</tr>
<tr>
<td>DIRAC</td>
<td>Is access to the quarters direct or through another unit? (1=direct, 2=another)</td>
<td>0.04</td>
</tr>
<tr>
<td>DIV</td>
<td>Census division (1=New England, 2=Middle Atlantic, 3=East North Central, 4=West North Central, 5=South Atlantic, 6=East South Central, 7=West South Central, 8=Mountain, 9=Pacific)</td>
<td>0.07</td>
</tr>
<tr>
<td>DOMSRVCQ</td>
<td>Domestic services this quarter</td>
<td></td>
</tr>
<tr>
<td>DMSXCCCQ</td>
<td>Domestic services excluding child care this quarter</td>
<td></td>
</tr>
<tr>
<td>DMSXCCPQ</td>
<td>Domestic services excluding child care last quarter</td>
<td></td>
</tr>
<tr>
<td>DOMSRVPQ</td>
<td>Domestic services last quarter</td>
<td></td>
</tr>
</tbody>
</table>

Wei-Yin Loh 329 GUIDE manual
<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EARNCOMP</td>
<td>Composition of earners (1=reference person only, 2=reference person and spouse, 3=reference person, spouse and others, 4=reference person and others, 5=spouse only, 6=spouse and others, 7=others, 8=no earners)</td>
</tr>
<tr>
<td>ECARTKNC</td>
<td>Outlays for new vehicle purchases this quarter including down payment, principal and interest paid on loans, or if not financed, purchase amount</td>
</tr>
<tr>
<td>ECARTKNP</td>
<td>Outlays for new vehicle purchases last quarter including down payment, principal and interest paid on loans, or if not financed, purchase amount</td>
</tr>
<tr>
<td>ECARTKUC</td>
<td>Outlays for used vehicle purchases this quarter including down payment, principal and interest paid on loans, or if not financed, purchase amount</td>
</tr>
<tr>
<td>ECARTKUP</td>
<td>Outlays for used vehicle purchases last quarter including down payment, principal and interest paid on loans, or if not financed, purchase amount</td>
</tr>
<tr>
<td>EDUC_REF</td>
<td>Education of reference person (10=grades 1–8; 11=grades 9–12, no degree; 12=high school graduate; 13=some college, no degree; 14=Associate’s degree in college; Bachelors degree; 16=Masters degree or professional/doctorate degree)</td>
</tr>
<tr>
<td>EDUCA2</td>
<td>Education level of spouse (same levels as EDUC_REF)</td>
</tr>
<tr>
<td>EDUCACQ</td>
<td>Education this quarter</td>
</tr>
<tr>
<td>EDUCAPQ</td>
<td>Education last quarter</td>
</tr>
<tr>
<td>EENTMSCC</td>
<td>Miscellaneous entertainment outlays this quarter including photographic and sports equipment and boat and RV rentals</td>
</tr>
<tr>
<td>EENTMSCP</td>
<td>Miscellaneous entertainment outlays last quarter including photographic and sports equipment and boat and RV rentals</td>
</tr>
<tr>
<td>EENTRMTC</td>
<td>Total entertainment outlays this quarter including sound systems, sports equipment, toys, cameras, and down payments on boats and campers (FEEADMCQ + TVR-DIOCQ + PETTOYCQ + EOTHENTC)</td>
</tr>
<tr>
<td>EENTRMTP</td>
<td>Total entertainment outlays last quarter including sound systems, sports equipment, toys, cameras, and down payments on boats and campers (same composition as EENTRMTC)</td>
</tr>
</tbody>
</table>

Wei-Yin Loh 330 GUIDE manual
<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EHOUSNGC</td>
<td>Total housing outlays this quarter including maintenance, fuels, public services, household operations, house furnishings, and mortgage (lump sum home equity loan or line of credit home equity loan) principle and interest (ESHELTRC + UTILCQ + HOUSOPCQ + HOUSEQCQ)</td>
</tr>
<tr>
<td>EHOUSNGP</td>
<td>Total housing outlays last quarter including maintenance, fuels, public services, household operations, house furnishings, and mortgage (lump sum home equity loan or line of credit home equity loan) principle and interest (same composition as EHOUSNGC)</td>
</tr>
<tr>
<td>ELCTRCCQ</td>
<td>Electricity this quarter</td>
</tr>
<tr>
<td>ELCTRCPQ</td>
<td>Electricity last quarter</td>
</tr>
<tr>
<td>EMISCELC</td>
<td>Miscellaneous outlays this quarter including reduction of mortgage principal (lump sum home equity loan) on other property (MISCPQ + EMISCMTP)</td>
</tr>
<tr>
<td>EMISCELP</td>
<td>Miscellaneous outlays last quarter including reduction of mortgage principal (lump sum home equity loan) on other property (same composition as EMISCELC)</td>
</tr>
<tr>
<td>EMISCMTTC</td>
<td>Mortgage principal outlays this quarter for other property</td>
</tr>
<tr>
<td>EMISCMTTP</td>
<td>Mortgage principal outlays last quarter for other property</td>
</tr>
<tr>
<td>EMRTPNOC</td>
<td>Mortgage principal outlays this quarter for owned home</td>
</tr>
<tr>
<td>EMRTPNOP</td>
<td>Mortgage principal outlays last quarter for owned home</td>
</tr>
<tr>
<td>EMRTPNVC</td>
<td>Mortgage principal outlays this quarter for owned vacation home</td>
</tr>
<tr>
<td>EMRTPNVP</td>
<td>Mortgage principal outlays last quarter for owned vacation home</td>
</tr>
<tr>
<td>EMOTRVHC</td>
<td>Outlays for motored recreational vehicles this quarter</td>
</tr>
<tr>
<td>EMOTRVHP</td>
<td>Outlays for motored recreational vehicles last quarter</td>
</tr>
<tr>
<td>ENOMOTRC</td>
<td>Outlays for non-motored recreational vehicles this quarter</td>
</tr>
<tr>
<td>ENOMOTRP</td>
<td>Outlays for non-motored recreational vehicles last quarter</td>
</tr>
<tr>
<td>ENTERTCQ</td>
<td>Entertainment this quarter (FEEADMCQ + TVRDIOCQ + OTHEQPCQ)</td>
</tr>
</tbody>
</table>
### CE VARIABLES

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENTERTPQ</td>
<td>Entertainment last quarter (same composition as ENTERTCQ)</td>
</tr>
<tr>
<td>EOTHENTC</td>
<td>Outlays for other entertainment supplies this quarter, equipment, and services including down payments on boats and campers (ENOMOTRC + EMOTRVHC + EENTMSCC)</td>
</tr>
<tr>
<td>EOTHENTP</td>
<td>Outlays for other entertainment supplies last quarter, equipment, and services including down payments on boats and campers (same composition as EOTHENTC)</td>
</tr>
<tr>
<td>EOTHLODC</td>
<td>Outlays for other lodging this quarter such as owned vacation home, including mortgage principal and interest, property taxes, maintenance, insurance, and other expenses (OTHLODCQ + EMRTPNVC)</td>
</tr>
<tr>
<td>EOTHLODP</td>
<td>Outlays for other lodging last quarter such as owned vacation home, including mortgage principal and interest, property taxes, maintenance, insurance, and other expenses (same composition as EOTHLODC)</td>
</tr>
<tr>
<td>EOTHVEHP</td>
<td>Outlays for other vehicle purchases last quarter such as motorcycles and airplanes including down payment, principal and interest paid on loans, or if not financed, purchase amount</td>
</tr>
<tr>
<td>EOTHVEHC</td>
<td>Outlays for other vehicle purchases this quarter such as motorcycles and airplanes including down payment, principal and interest paid on loans, or if not financed, purchase amount</td>
</tr>
<tr>
<td>EOTHVEHP</td>
<td>Outlays for other vehicle purchases last quarter such as motorcycles and airplanes including down payment, principal and interest paid on loans, or if not financed, purchase amount</td>
</tr>
<tr>
<td>EOWNDWLC</td>
<td>Owned home outlays this quarter including mortgage principal and interest, property taxes, maintenance, insurance, and other expenses (OWNDWECQ + EMRTPNOC)</td>
</tr>
<tr>
<td>EOWNDWLP</td>
<td>Owned home outlays last quarter including mortgage principal and interest, property taxes, maintenance, insurance, and other expenses (same composition as EOWNDWLC)</td>
</tr>
<tr>
<td>ERANKH</td>
<td>Percent expenditure outlay rank</td>
</tr>
<tr>
<td>Variable</td>
<td>Description</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>ERANKHM</td>
<td>Weighted cumulative percent expenditure outlay ranking of CU to total population</td>
</tr>
<tr>
<td>ESHELTRC</td>
<td>Shelter outlays this quarter including mortgage principle and interest for owned home and/or vacation home, rents, insurance, taxes, and maintenance (EOWNDWLC + RENDWECQ + EOTHLODC)</td>
</tr>
<tr>
<td>ESHELTRP</td>
<td>Shelter outlays last quarter including mortgage principle and interest for owned home and/or vacation home, rents, insurance, taxes, and maintenance (same composition as ESHELTRC)</td>
</tr>
<tr>
<td>ETOTALC</td>
<td>Total outlays this quarter, sum of outlays from all major expenditure categories (FOODCQ + ALCBEVCQ + EHOUSNGC + APPARCQ + ETRANPTC + HEALTHCQ + EENTRMTC + PERSCACQ + READCQ + EDUCACQ + TOBACCCQ + EMISCELC + CASHCOCQ + PERINSCQ)</td>
</tr>
<tr>
<td>ETOTALP</td>
<td>Total outlays last quarter, sum of outlays from all major expenditure categories (same composition as ETOTALC)</td>
</tr>
<tr>
<td>ETOTACX4</td>
<td>Adjusted total outlays this quarter, sum of outlays from all major expenditure categories (FOODCQ + ALCBEVCQ + EHOUSNGC + APPARCQ + ETRANPTC + HEALTHCQ + EENTRMTC + PERSCACQ + READCQ + EDUCACQ + TOBACCCQ + MISC1CQ + 4 × MISC2CQ + EMISCMTTC + PERINSCQ)</td>
</tr>
<tr>
<td>ETOTAPX4</td>
<td>Adjusted total outlays last quarter, sum of outlays from all major expenditure categories (same composition as ETOTACX4)</td>
</tr>
<tr>
<td>ETRANPTC</td>
<td>Total outlays for transportation this quarter including down payment, principal and finance charges paid on loans, gasoline and motor oil, maintenance and repairs, insurance, public and other transportation, and vehicle rental licenses and other charges (EVEHPURC + GASMOCQ + MAINRPCQ + VEHINSCQ + VRNTLOCQ + PUBTRACQ)</td>
</tr>
</tbody>
</table>

Wei-Yin Loh

333

GUIDE manual
ETRANPTP: Total outlays for transportation last quarter including down payment, principal and finance charges paid on loans, gasoline and motor oil, maintenance and repairs, insurance, public and other transportation, and vehicle rental licenses and other charges (same composition as ETRANPTC).

EVEHPURC: Outlays for vehicle purchases this quarter including down payment, principal and interest paid on loans, or if not financed, purchase amount (ECARTKNC + ECARTKUC + EOTHVEHC).

EVEHPURP: Outlays for vehicle purchases last quarter including down payment, principal and interest paid on loans, or if not financed, purchase amount (same composition as EVEH PURC).

FAM_SIZE: Number of Members in CU.

FAM_TYPE: Family type (1–9).

FDAWAYCQ: Food away from home this quarter.

FDAWAYPQ: Food away from home last quarter.

FDHOMEPCQ: Food at home this quarter.

FDHOMEQPQ: Food at home last quarter.

FDMAPCQ: Meals as pay this quarter.

FDMAPPQ: Meals as pay last quarter.

FDXMAPCQ: Food away excluding meals as pay this quarter.

FDXMAPPQ: Food away excluding meals as pay last quarter.

FEEADMCQ: Fees and admissions this quarter.

FEEADMPQ: Fees and admissions last quarter.

FFTAXOWE: Weighted estimate for Federal tax liabilities for entire CU.

FGOVRETM: Amount of government retirement deducted from last pay, annualized for all CU members.

FGOVRETX: Amount of government retirement deducted from last pay annualized.

FINCBTAX: Total family income before taxes in last 12 months (INTRDVX + INTRDVBX + ROYESTX + ROYESTBX + OTHREGX + OTHREGBX + WELFAREX + WEFREBX + RETSURVX + RETSRVBX + NETRENTX + NETRNTBX + OTHRINCX).

FINDRETX: Money placed in self-employed retirement plan in past year for all CU members.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FINLWT21</td>
<td>Sampling weight</td>
</tr>
<tr>
<td>FJSSDEDX</td>
<td>Estimated amount contributed to Social Security by all CU members past 12 mos.</td>
</tr>
<tr>
<td>FLRCVRCQ</td>
<td>Floor coverings this quarter</td>
</tr>
<tr>
<td>FLRCVRPQ</td>
<td>Floor coverings last quarter</td>
</tr>
<tr>
<td>FMLPYYRX</td>
<td>Annual value of free meals received as part of pay</td>
</tr>
<tr>
<td>FOODCQ</td>
<td>Total food this quarter</td>
</tr>
<tr>
<td>FOODPQ</td>
<td>Total food last quarter</td>
</tr>
<tr>
<td>FPRIPENM</td>
<td>Amount of private pensions deducted from last pay, annualized, for all CU members</td>
</tr>
<tr>
<td>FPRIPENX</td>
<td>Amount of private pensions</td>
</tr>
<tr>
<td>FRRDEDM</td>
<td>Amount of Railroad Retirement deducted from last pay, annualized for all CU members</td>
</tr>
<tr>
<td>FRRDEDX</td>
<td>Amount of railroad retirement deducted from last pay, annualized</td>
</tr>
<tr>
<td>FRRETIRX</td>
<td>Social security and railroad retirement income</td>
</tr>
<tr>
<td>FS_MTHI</td>
<td>In how many of the last 12 months were food stamps or EBTs received?</td>
</tr>
<tr>
<td>FSALARXY</td>
<td>Wage and salary income of all members past 12 mos.</td>
</tr>
<tr>
<td>FSMPFRMX</td>
<td>Family level summation for new variable SEMPFRMX and SMPFRMBX</td>
</tr>
<tr>
<td>FSSIX</td>
<td>Amount supplemental security income from all sources received by all CU members in past 12 months</td>
</tr>
<tr>
<td>FSTAXOWE</td>
<td>Weighted estimate for State tax liabilities for entire CU</td>
</tr>
<tr>
<td>FURNTRCQ</td>
<td>Furniture this quarter</td>
</tr>
<tr>
<td>FURNTRPQ</td>
<td>Furniture last quarter</td>
</tr>
<tr>
<td>GASMOCQ</td>
<td>Gasoline and motor oil this quarter</td>
</tr>
<tr>
<td>GASMOPQ</td>
<td>Gasoline and motor oil last quarter</td>
</tr>
<tr>
<td>FULOILCQ</td>
<td>Fuel oil this quarter</td>
</tr>
<tr>
<td>FULOILPQ</td>
<td>Fuel oil last quarter</td>
</tr>
<tr>
<td>FURNTRCQ</td>
<td>Furniture this quarter</td>
</tr>
<tr>
<td>FURNTRPQ</td>
<td>Furniture last quarter</td>
</tr>
<tr>
<td>HEALTHCQ</td>
<td>Health care this quarter (HLTHINCQ + MEDSRVCQ + PREDRGCQ + MEDSUPCQ)</td>
</tr>
</tbody>
</table>

Wei-Yin Loh  335  GUIDE manual
HEALTHPQ  Health care last quarter (same composition as HEALTHCQ)
HIGH_EDU  Highest level of education within the CU (0=never attended, 10=1–8 grade, 11=9–12 grade, 12=HS grad, 13=some college, 14=AA degree, 15=Bachelors, 16=Masters/professional/doctorate)
HISP_REF  Hispanic origin of reference person (1=Hispanic, 2=non-Hispanic)
HISP2  Hispanic origin of spouse (1=Hispanic, 2=non-Hispanic)
HH_CU_Q  Count of CUs in household
HLFBATHQ  How many half bathrooms are there in this unit? 0.01
HLTHINCQ  Health insurance this quarter
HLTHINPQ  Health insurance last quarter
HORREF1  Hispanic origin of reference person (1=Mexican, 2=Mexican-American, 3=Chicano, 4=Puerto Rican, 5=Cuban, 6=Other) 0.96
HORREF2  Hispanic origin of spouse (same codes as HORREF1) 0.98
HOUSCQ  Housing this quarter
HOUSEQCQ  House furnishings and equipment this quarter (TEXTILCQ + FURNTRCQ + FLRCVRCQ + MAJAPPCQ + SMLAPPCQ + MISCEQCQ)
HOUSEQPQ  House furnishings and equipment last quarter (same composition as HOUSEQCQ)
HOUSOPCQ  Household operations this quarter
HOUSPQ  Housing last quarter
HOUSOPPQ  Household operations last quarter
INC_HRS1  Number hours worked per week by reference person 0.38
INC_HRS2  Number hours worked per week by spouse 0.66
INC_RANK  Income rank of CU to total population
INCLASS2  Income class based on INC_RANK (1=0–0.1667, 2=0.1667–0.3333, 3=0.3334–0.4999, 4=0.5000–0.6666, 5=0.6667–0.8333, 6=0.8334–1, 7=incomplete reporting)
INCNONW1  Reason for not working during past 12 months (1=retired, 2=take care of home, 3=going to school, 4=ill, disabled, unable to work, 5=unable to find work, 6=doing something else) 0.62
INCNONW2  Reason spouse did not work during past 12 months (same codes as INCNONW1) 0.78
<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>INCOMEY1</td>
<td>Employer paying most earnings in past 12 months (1=private company, business or individual, 2=Federal govt, 3=State govt, 4=local govt, 5=self-employed, 6=family business or farm, working without pay)</td>
<td>0.38</td>
</tr>
<tr>
<td>INCOMEY2</td>
<td>Employer from which spouse received most earnings during the past 12 months</td>
<td>0.66</td>
</tr>
<tr>
<td>INCWEEK1</td>
<td>Weeks worked full or part time in last 12 months</td>
<td>0.44</td>
</tr>
<tr>
<td>INCWEEK2</td>
<td>Weeks worked by spouse full or part time last 12 months</td>
<td>0.44</td>
</tr>
<tr>
<td>INTRDVX</td>
<td>Amount received in interest or dividend during past 12 mos.</td>
<td>0.37</td>
</tr>
<tr>
<td>IRA</td>
<td>Do you have any retirement accounts such as 401(k)s, IRAs, thrift saving plans? (1=yes, 2=no)</td>
<td>0.76</td>
</tr>
<tr>
<td>IRAB</td>
<td>Range that best reflects the total value of all retirement accounts such as 401(k)s, IRAs, and thrift savings plans (1=0–1999, 2=2000-9999, 3=10K–49999, 4=50K–199999, 5=200K–449999, 6=450K or more)</td>
<td>0.97</td>
</tr>
<tr>
<td>IRAX</td>
<td>Total amount put into retirement accounts past 12 mos.</td>
<td>0.87</td>
</tr>
<tr>
<td>IRAYRB</td>
<td>Range which best reflects the total value of all retirement accounts one year ago today (same codes as IRAB)</td>
<td>0.96</td>
</tr>
<tr>
<td>IRAYRBX</td>
<td>Median value of bracket range for IRAYRB</td>
<td>0.96</td>
</tr>
<tr>
<td>IRAYRX</td>
<td>Total value of retirement accounts one year ago</td>
<td>0.88</td>
</tr>
<tr>
<td>JFS_AMT</td>
<td>Annual value of food stamps</td>
<td></td>
</tr>
<tr>
<td>LIFINSCQ</td>
<td>Life and other personal insurance this quarter</td>
<td></td>
</tr>
<tr>
<td>LIFINSPQ</td>
<td>Life and other personal insurance last quarter</td>
<td></td>
</tr>
<tr>
<td>LIQUDYRBX</td>
<td>Median value of bracket range for LIQUDYRB</td>
<td>0.96</td>
</tr>
<tr>
<td>LIQUDYR</td>
<td>Did you have any checking savings money market accounts, or CDs one year ago? (1=yes, 2=no)</td>
<td>&gt;0.99</td>
</tr>
<tr>
<td>LIQUID</td>
<td>Do you have any checking, saving, money market accounts, or CDs? (1=yes, 2=no)</td>
<td>0.76</td>
</tr>
<tr>
<td>LIQUIDB</td>
<td>Range that best reflects total value of checking, savings, money market accounts, CDs (1=0–499, 2=500-999, 3=1000–2499, 4=2.5K–99999, 5=10K–349999, 6=35K and over)</td>
<td>0.97</td>
</tr>
<tr>
<td>LIQUIDBX</td>
<td>Median value of bracket range LIQUIDB</td>
<td>0.97</td>
</tr>
<tr>
<td>LIQUIDX</td>
<td>Total value of all checking, savings, money market, and CD accounts</td>
<td>0.83</td>
</tr>
<tr>
<td>Variable</td>
<td>Description</td>
<td>Code</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>LQUDYRB</td>
<td>Range that best reflects the total value of all checking, savings, money market accounts, and CDs one year ago today (same codes as LIQUIDB)</td>
<td>0.97</td>
</tr>
<tr>
<td>LQUDYRX</td>
<td>Total value of all checking, savings, money market accounts, and CDs one year ago today</td>
<td>0.84</td>
</tr>
<tr>
<td>MAINRPCQ</td>
<td>Maintenance and repairs this quarter</td>
<td></td>
</tr>
<tr>
<td>MAINRPPQ</td>
<td>Maintenance and repairs last quarter</td>
<td></td>
</tr>
<tr>
<td>MAJAPPCQ</td>
<td>Major appliances this quarter</td>
<td></td>
</tr>
<tr>
<td>MAJAPPPQ</td>
<td>Major appliances last quarter</td>
<td></td>
</tr>
<tr>
<td>MARITAL1</td>
<td>Marital status of reference person (1=married, 2=widowed, 3=divorced, 4=separated, 5=never married)</td>
<td></td>
</tr>
<tr>
<td>MEALSPAY</td>
<td>Have you received any free meals at work as part of your pay? (1=yes, 2=no)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>MEDSRVCQ</td>
<td>Medical services this quarter</td>
<td></td>
</tr>
<tr>
<td>MEDSRVPQ</td>
<td>Medical services last quarter</td>
<td></td>
</tr>
<tr>
<td>MEDSUPCQ</td>
<td>Medical supplies this quarter</td>
<td></td>
</tr>
<tr>
<td>MEDSUPPQ</td>
<td>Medical supplies last quarter</td>
<td></td>
</tr>
<tr>
<td>MENBOYCQ</td>
<td>Clothing for men and boys this quarter</td>
<td></td>
</tr>
<tr>
<td>MENSIXCQ</td>
<td>Clothing for men, 16 and over this quarter</td>
<td></td>
</tr>
<tr>
<td>MENSIXPQ</td>
<td>Clothing for men, 16 and over last quarter</td>
<td></td>
</tr>
<tr>
<td>MENBOYPQ</td>
<td>Clothing for men and boys last quarter</td>
<td></td>
</tr>
<tr>
<td>MISC1CQ</td>
<td>Miscellaneous expenditures this quarter</td>
<td></td>
</tr>
<tr>
<td>MISC1PQ</td>
<td>Miscellaneous expenditures last quarter</td>
<td></td>
</tr>
<tr>
<td>MISCEQCQ</td>
<td>Miscellaneous household equipment this quarter</td>
<td></td>
</tr>
<tr>
<td>MISCEQPQ</td>
<td>Miscellaneous household equipment last quarter</td>
<td></td>
</tr>
<tr>
<td>MISCSCQ</td>
<td>Miscellaneous expenditures this quarter (MISC1CQ + MISC2CQ)</td>
<td></td>
</tr>
<tr>
<td>MISCSPQ</td>
<td>Miscellaneous expenditures last quarter (same composition as MISCSCQ)</td>
<td></td>
</tr>
<tr>
<td>MISCTXAX</td>
<td>During past 12 months, what was total amount paid for personal property taxes and other taxes not reported elsewhere by all CU members?</td>
<td>0.99</td>
</tr>
<tr>
<td>MISCX4CQ</td>
<td>Adjusted miscellaneous expenditures this quarter (MISC1CQ + 4×MISC2CQ)</td>
<td></td>
</tr>
<tr>
<td>MISCX4PQ</td>
<td>Adjusted miscellaneous expenditures last quarter (same composition as MISCX4CQ)</td>
<td></td>
</tr>
<tr>
<td>MLPAYWKX</td>
<td>About what was the weekly dollar value of these meals?</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>Description</td>
<td>Code</td>
</tr>
<tr>
<td>---</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>MLPYQWKS</td>
<td>For how many weeks did members of your household receive these meals during the past 12 months?</td>
<td>0.99</td>
</tr>
<tr>
<td>MRPINSCQ</td>
<td>Maintenance, repairs, insurance, and other expenses this quarter</td>
<td></td>
</tr>
<tr>
<td>MRPINSPQ</td>
<td>Maintenance, repairs, insurance, and other expenses last quarter</td>
<td></td>
</tr>
<tr>
<td>MRTINTCQ</td>
<td>Mortgage interest this quarter</td>
<td></td>
</tr>
<tr>
<td>MRTINTPQ</td>
<td>Mortgage interest last quarter</td>
<td></td>
</tr>
<tr>
<td>MRTPRNOC</td>
<td>Outlays on owned vacation home mortgage principle this quarter</td>
<td></td>
</tr>
<tr>
<td>MRTPRNOP</td>
<td>Outlays on owned vacation home mortgage principle last quarter</td>
<td></td>
</tr>
<tr>
<td>NETRENTB</td>
<td>Range that best reflects the total net rental income or loss during the past 12 months (1=0–999, 2=1–2K, 3=2–3K, 4=3–4K, 5=4–5K, 6=5–10K, 7=10–15K, 8=15–20K, 9=20–30K, 10=30–40K, 11=40–50K, 12=50K and over)</td>
<td>0.99</td>
</tr>
<tr>
<td>NETRENTX</td>
<td>What was the amount of net rental income or loss?</td>
<td>0.92</td>
</tr>
<tr>
<td>NETRENTBX</td>
<td>Median value of bracket range of NETRENTB</td>
<td>0.99</td>
</tr>
<tr>
<td>NTLGASCQ</td>
<td>Natural gas this quarter</td>
<td></td>
</tr>
<tr>
<td>NTLGASPQ</td>
<td>Natural gas last quarter</td>
<td></td>
</tr>
<tr>
<td>NO_EARNR</td>
<td>Number of earners</td>
<td></td>
</tr>
<tr>
<td>NONINCMX</td>
<td>Amount of other money receipts excluded from CU income before taxes received by CU in past 12 months</td>
<td></td>
</tr>
<tr>
<td>NUM_AUTO</td>
<td>Total number of owned cars</td>
<td></td>
</tr>
<tr>
<td>NUM_TVAN</td>
<td>Total number of owned trucks and vans</td>
<td></td>
</tr>
<tr>
<td>OCCUCOD1</td>
<td>Highest paid occupation last 12 months (15 coded values)</td>
<td>0.38</td>
</tr>
<tr>
<td>OCCUCOD1</td>
<td>Job in which reference person received most earnings during past 12 months (15 coded values)</td>
<td>0.66</td>
</tr>
<tr>
<td>OCCUCOD2</td>
<td>Job in which spouse received most earnings during past 12 months (15 coded values)</td>
<td>0.66</td>
</tr>
<tr>
<td>OTHAPLCQ</td>
<td>Other apparel products and services this quarter</td>
<td></td>
</tr>
<tr>
<td>OTHAPLPQ</td>
<td>Other apparel products and services last quarter</td>
<td></td>
</tr>
<tr>
<td>OTHASTB</td>
<td>Range which best reflects the total value of these other financial assets (1=0–2K, 2=2–10K, 3=10–50K, 4=50–200K, 5=200–450K, 6=450K and over)</td>
<td>&gt;0.99</td>
</tr>
<tr>
<td>OTHASTBX</td>
<td>Median value of bracket range for OTHASTB</td>
<td>&gt;0.99</td>
</tr>
<tr>
<td>OTHASTX</td>
<td>Total value of these other financial assets as of today</td>
<td>0.99</td>
</tr>
<tr>
<td>OTHENTCQ</td>
<td>Other entertainment this quarter</td>
<td></td>
</tr>
</tbody>
</table>
A CE VARIABLES

OTHENTPQ Other entertainment last quarter
OTHQPCQ Other equipment and services this quarter (PETTOYCQ + OTHENTCQ)
OTHQPPQ Other equipment and services last quarter (same composition as OTHQPCQ)
OTHFINX Total amount paid in finance, late charges, and interest for all other loans in the last month 0.99
OTHFLSCQ Other fuels this quarter
OTHFLSPQ Other fuels last quarter
OTHHEXCQ Other household expenses this quarter
OTHHEXPQ Other household expenses last quarter
OTHLNYR Did you have any other debt such as medical loans or personal loans one year ago today? (1=yes, 2=no) >0.99
OTHLNYRB Range which best reflects the total amount owed on all other loans one year ago today (1=0–499, 2=500-999, 3=1–2.5K, 4=2.5–10K, 5=10–35K, 6=35K and over) >0.99
OTHLODCQ Other lodging this quarter
OTHLODPQ Other lodging last quarter
OTHLONX Total amount owed on all other loans 0.99
OTHLYRBX Median value of bracket range for OTHLONBX >0.99
OTHREGB Range best reflects total amount received in Veteran’s Administration (VA) payments, unemployment compensation, child support, or alimony during the past 12 months (1=0–1K, 2=1–2K, 3=2–3K, 4=3–4K, 5=4–5K, 6=5–10K, 7=10–15K, 8=15–20K, 9=20–30K, 10=30–40K, 11=40–50K, 12=50K and over) 0.99
OTHREGBX Median value of bracket range for OTHREGB 0.99
OTHRINCX Amount received in other income including money from care of foster children, cash scholarships and fellowships, or stipends not based on working 0.97
OTHRREGX Income on a regular basis from any other source such as Veteran’s Administration (VA) payments, unemployment compensation, child support, or alimony 0.92
OTHSTYRB Range which best reflects total value of these other financial assets one year ago today (1=0–2K, 2=2–10K, 3=10–50K, 4=50–200K, 5=200–450K, 6=450K and over) >0.99
OTHSTYRX Value of these other financial assets one year ago today 0.99
OTHSYRBX Median value of bracket range for OTHSTYRB >0.99
A  CE VARIABLES

OTHVEHCQ  Other vehicles this quarter
OTHVEHPQ  Other vehicles last quarter
OWNDWECQ  Owned dwellings this quarter (MRTINTCQ + PROPTXCQ + MRPINSCQ)
OWNDWEPQ  Owned dwellings last quarter (same composition as OWNDWECQ)
OWNVACC  Expenditures on owned vacation homes this quarter including mortgage interest, insurance, taxes, maintenance, and miscellaneous household equipment (VOTHRLOC + VMISCHEC)
OWNVACP  Expenditures on owned vacation homes last quarter including mortgage interest, insurance, taxes, maintenance, and miscellaneous household equipment (same composition as OWNVACC)
PERINSCQ  Personal insurance and pensions this quarter (LIFINSCQ + RETPENCQ)
PERINSPQ  Personal insurance and pensions last quarter (same composition as PERINSCQ)
PERSCACQ  Personal care this quarter
PERSCAPQ  Personal care last quarter
PERSLT18  Number of CU members less than 18
PERSOT64  Number of CU members over 64
PETTOYCQ  Pets, toys, and playground equipment this quarter
PETTOYPQ  Pets, toys, and playground equipment last quarter
POPSIZE  Population size of the PSU (1=more than 5M, 2=1–5M, 3=0.5–1M, 4=100–500K, 5=less than 100K)
PREDRCGCQ  Prescription drugs this quarter
PREDRGPQ  Prescription drugs last quarter
PRINEARN  Member number of principal earner (5 coded values)
PROPTXCQ  Property taxes this quarter
PROPTXPQ  Property taxes last quarter
PSU  Primary sampling unit
0.52
PUBTRACQ  Public and other transportation this quarter (TRNTRPCQ + TRNOTHCQ)
PUBTRAPQ  Public and other transportation last quarter (same composition as PUBTRACQ)
RACE2  Race of spouse (same codes as REF_RACE)
0.44
READCQ  Reading this quarter

Wei-Yin Loh  341  GUIDE manual
A CE VARIABLES

READPQ Reading last quarter
REF_RACE Race of reference person (1=white, 2=black, 3=native American, 4=Asian, 5=Pacific islander, 6=multi-race)
REGION Region (1=Northeast, 2=Midwest, 3=South, 4=West)
RELECTRC Expenditures on electricity for rented vacation homes this quarter
RELECTRTP Expenditures on electricity for rented vacation homes last quarter
RENDWECQ Rented dwelling this quarter (RNTXRPCQ + RNTAPYPCQ)
RENDWEPQ Rented dwelling last quarter (same composition as RENDWECQ)
RENTEQVX Monthly rent if home rented today
RETPENCQ Retirement, pensions, social security this quarter
RETPENPQ Retirement, pensions, social security last quarter
RETSRVBX Median value of bracket range for RETSURVB
RETSURV Did you receive income from retirement, survivor, or disability pensions during past 12 months? (1=yes, 2=no)
RETSURVX Retirement, survivor, disability pensions received past 12 mos.
RNA TLGAC Expenditures on natural gas for rented vacation homes this quarter
RNA TLGAP Expenditures on natural gas for rented vacation homes last quarter
RNTAPYPCQ Rent as pay this quarter
RNTAPYPQ Rent as pay last quarter
RNTXRPCQ Rent excluding rent as pay this quarter
RNTXRPPQ Rent excluding rent as pay last quarter
ROOMSQ Number of rooms in CU living quarters, including finished living areas, excluding all baths
ROTHRFLC Expenditures on other fuels for rented vacation homes this quarter
<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROYESTB</td>
<td>Range that best reflects total amount received in royalty income or income from estates and trusts during past 12 months (1=0–1K, 2=1–2K, 3=2–3K, 4=3–4K, 5=4–5K, 6=5–10K, 7=10–15K, 8=15–20K, 9=20–30K, 10=30–40K, 11=40–50K, 12=50K and over)</td>
<td>&gt;0.99</td>
</tr>
<tr>
<td>ROYESTBX</td>
<td>Median value of bracket range for ROYESTB</td>
<td>&gt;0.99</td>
</tr>
<tr>
<td>ROYESTX</td>
<td>Amount received in royalty income or income from estates and trusts</td>
<td>0.96</td>
</tr>
<tr>
<td>RWATERPC</td>
<td>Expenditures on water and public services for rented vacation homes this quarter</td>
<td></td>
</tr>
<tr>
<td>RWATERPP</td>
<td>Expenditures on water and public services for rented vacation homes last quarter</td>
<td></td>
</tr>
<tr>
<td>SEX_REF</td>
<td>Sex of reference person (1=male, 2=female)</td>
<td></td>
</tr>
<tr>
<td>SEX2</td>
<td>Sex of spouse (1=male, 2=female)</td>
<td>0.44</td>
</tr>
<tr>
<td>SHELTCQ</td>
<td>Shelter this quarter (OWNDWECQ + RENDWECQ + OTHLODCQ)</td>
<td></td>
</tr>
<tr>
<td>SHELTPQ</td>
<td>Shelter last quarter (same composition as SHELTCQ)</td>
<td></td>
</tr>
<tr>
<td>SMLAPPCQ</td>
<td>Small appliances, miscellaneous housewares this quarter</td>
<td></td>
</tr>
<tr>
<td>SMLAPPPQ</td>
<td>Small appliances, miscellaneous housewares last quarter</td>
<td></td>
</tr>
<tr>
<td>SMSASTAT</td>
<td>Does CU reside inside a Metropolitan Statistical Area (MSA)? (1=yes, 2=no)</td>
<td></td>
</tr>
<tr>
<td>ST_HOUS</td>
<td>Are these living quarters presently used as student housing by a college or university? (1=yes, 2=no)</td>
<td></td>
</tr>
<tr>
<td>STATE</td>
<td>1=AL, 2=AK, 4=AZ, 5=AR, 6=CA, 8=CO, 9=CT, 10=DE, 11=DC, 12=FL, 13=GA, 15=HI, 16=ID, 17=IL, 18=IN, 19=IA, 20=KS, 21=KY, 22=LA, 23=ME, 24=MD, 25=MA, 26=MI, 27=MN, 28=MS, 29=MO, 30=MT, 31=NE, 32=NV, 33=NH, 34=NJ, 36=NY, 37=NC, 39=OH, 40=OK, 41=OR, 42=PA, 44=RI, 45=SC, 46=SD, 47=TN, 48=TX, 49=UT, 51=VA, 53=WA, 54=WV, 55=WI</td>
<td>0.08</td>
</tr>
<tr>
<td>STCKYRBX</td>
<td>Median value of bracket range for STOCKYRB</td>
<td>0.98</td>
</tr>
<tr>
<td>STDNTYR</td>
<td>Did you have student loans one years ago today? (1=yes, 2=no)</td>
<td>&gt;0.99</td>
</tr>
<tr>
<td>STDNTYRB</td>
<td>Range which best reflects the total amount owed on all student loans one year ago today (1=0–499, 2=500–999, 3=1–2.5K, 4=2.5–10K, 5=10–35K, 6=35K and over)</td>
<td>&gt;0.99</td>
</tr>
<tr>
<td>Variable</td>
<td>Description</td>
<td>Value</td>
</tr>
<tr>
<td>---------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>STDNTYRX</td>
<td>Total amount owed on all student loans one year ago today</td>
<td>0.97</td>
</tr>
<tr>
<td>STDTRYRBX</td>
<td>Median value of bracket range for STDNTYRB</td>
<td>&gt;0.99</td>
</tr>
<tr>
<td>STOCKB</td>
<td>Range which best reflects total value of all directly-held stocks, bonds, and mutual funds (1=0–2K, 2=2–10K, 3=10–50K, 4=50–200K, 5=200–450K, 6=450K and over)</td>
<td>0.99</td>
</tr>
<tr>
<td>STOCKBX</td>
<td>Median value of bracket range for STOCKB</td>
<td>0.99</td>
</tr>
<tr>
<td>STOCKX</td>
<td>Value of directly-held stocks, bonds, mutual funds (median=59,950, mean=411,867)</td>
<td>0.93</td>
</tr>
<tr>
<td>STOCKYR</td>
<td>Did you have any directly-held stocks, bonds, or mutual funds one year ago? (1=yes, 2=no)</td>
<td>&gt;0.99</td>
</tr>
<tr>
<td>STOCKYRB</td>
<td>Range which best reflects total value of all directly-held stocks, bonds, and mutual funds one year ago today (same codes as STOCKB)</td>
<td>0.98</td>
</tr>
<tr>
<td>STOCKYRX</td>
<td>Median value of bracket range of STOCKX</td>
<td>0.93</td>
</tr>
<tr>
<td>STUDFINX</td>
<td>Total amount paid in finance, late charges, and interest for all student loans in the last month</td>
<td>0.97</td>
</tr>
<tr>
<td>STUDNTB</td>
<td>Range which best reflects the total amount owed on all student loans (1=0–499, 2=500–999, 3=1–2.5K, 4=2.5–10K, 5=10–35K, 6=35K and over)</td>
<td>&gt;0.99</td>
</tr>
<tr>
<td>STUDNTBX</td>
<td>Median value of bracket range for STUDNTB</td>
<td>&gt;0.99</td>
</tr>
<tr>
<td>STUDNTX</td>
<td>Total amount owed on all student loans</td>
<td>0.97</td>
</tr>
<tr>
<td>TAIRFARC</td>
<td>Trip expenditures on airfare this quarter</td>
<td></td>
</tr>
<tr>
<td>TAIRFARP</td>
<td>Trip expenditures on airfare last quarter</td>
<td></td>
</tr>
<tr>
<td>TALCBEVC</td>
<td>Total trip expenditures this quarter on alcoholic beverages at restaurants, cafes, and bars</td>
<td></td>
</tr>
<tr>
<td>TALCBEVP</td>
<td>Total trip expenditures last quarter on alcoholic beverages at restaurants, cafes, and bars</td>
<td></td>
</tr>
<tr>
<td>TELEPHCQ</td>
<td>Telephone services this quarter</td>
<td></td>
</tr>
<tr>
<td>TELEPHPQ</td>
<td>Telephone services last quarter</td>
<td></td>
</tr>
<tr>
<td>TENTRMNC</td>
<td>Total trip expenditures on entertainment this quarter including sporting events, movies, and recreational vehicle rentals (TFEESADC + TOTHENTC)</td>
<td></td>
</tr>
<tr>
<td>TENTRMNP</td>
<td>Total trip expenditures on entertainment last quarter including sporting events, movies, and recreational vehicle rentals (same composition as TENTRMNC)</td>
<td></td>
</tr>
<tr>
<td>TEXTILCQ</td>
<td>Household textiles this quarter</td>
<td></td>
</tr>
<tr>
<td>TEXTILPQ</td>
<td>Household textiles last quarter</td>
<td></td>
</tr>
<tr>
<td>Variable</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>TFAREC</td>
<td>Trip expenditures this quarter on transportation fares including airfare, intercity bus, train, and ship fare (TAIRFARC + TOTHFARC)</td>
<td></td>
</tr>
<tr>
<td>TFAREP</td>
<td>Trip expenditures last quarter on transportation fares including airfare, intercity bus, train, and ship fare (same composition as TFAREC)</td>
<td></td>
</tr>
<tr>
<td>TFEESADC</td>
<td>Trip expenditures on miscellaneous entertainment this quarter including recreation expenses, participation sport fees, and admission fees to sporting events and movies</td>
<td></td>
</tr>
<tr>
<td>TFEESADP</td>
<td>Trip expenditures on miscellaneous entertainment last quarter including recreation expenses, participation sport fees, and admission fees to sporting events and movies</td>
<td></td>
</tr>
<tr>
<td>TFOODAWC</td>
<td>Food and non-alcoholic beverages this quarter at restaurants, cafes, and fast food places during out-of-town trips</td>
<td></td>
</tr>
<tr>
<td>TFOODAWP</td>
<td>Food and non-alcoholic beverages last quarter at restaurants, cafes, and fast food places during out-of-town trips</td>
<td></td>
</tr>
<tr>
<td>TFOODHOC</td>
<td>Food and beverages purchased and prepared by CU this quarter during out-of-town trips</td>
<td></td>
</tr>
<tr>
<td>TFOODHOP</td>
<td>Food and beverages purchased and prepared by CU last quarter during out-of-town trips</td>
<td></td>
</tr>
<tr>
<td>TFOODTOC</td>
<td>Total trip expenditures on food this quarter including both restaurant food and food prepared by CU (TFOODAWC + TFOODHOC)</td>
<td></td>
</tr>
<tr>
<td>TFOODTOP</td>
<td>Total trip expenditures on food last quarter including both restaurant food and food prepared by CU (same composition as TFOODTOC)</td>
<td></td>
</tr>
<tr>
<td>TGASMO TC</td>
<td>Trip expenditures on gas and oil this quarter</td>
<td></td>
</tr>
<tr>
<td>TGASMO TP</td>
<td>Trip expenditures on gas and oil last quarter</td>
<td></td>
</tr>
<tr>
<td>TLOCALTC</td>
<td>Trip expenditures this quarter on local transportation including taxis, buses etc.</td>
<td></td>
</tr>
<tr>
<td>TLOCALTP</td>
<td>Trip expenditures last quarter on local transportation including taxis, buses etc.</td>
<td></td>
</tr>
<tr>
<td>TOBACCCQ</td>
<td>Tobacco and smoking supplies this quarter</td>
<td></td>
</tr>
<tr>
<td>TOBACCPQ</td>
<td>Tobacco and smoking supplies last quarter</td>
<td></td>
</tr>
<tr>
<td>TOTEX4CQ</td>
<td>Adjusted total expenditures this quarter (TOTEXPCQ - MISCCQ + MISC1CQ + 4 × MISC2CQ)</td>
<td></td>
</tr>
<tr>
<td>TOTEX4PQ</td>
<td>Adjusted total expenditures last quarter (same composition as TOTEX4CQ)</td>
<td></td>
</tr>
<tr>
<td>Variable</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>TOTEXPCQ</td>
<td>Total expenditures this quarter (FOODCQ + AL-CBEVCQ + HOUSCQ + APPARCQ + TRANSCQ + HEALTHCQ + ENTERTCQ + PERSCACQ + READCQ + EDUCACQ + TOBACCCQ + MISCCQ + CASHCOQCQ + PERINSCQCQ)</td>
<td></td>
</tr>
<tr>
<td>TOTEXPPQ</td>
<td>Total expenditures last quarter (same composition as TOTEXPCQ)</td>
<td></td>
</tr>
<tr>
<td>TOTHENTC</td>
<td>Trip expenditures on recreational vehicle rentals this quarter including campers, boats, and other vehicles</td>
<td></td>
</tr>
<tr>
<td>TOTHENTP</td>
<td>Trip expenditures on recreational vehicle rentals last quarter including campers, boats, and other vehicles</td>
<td></td>
</tr>
<tr>
<td>TOTHFARC</td>
<td>Tip expenditures this quarter on other transportation fares including intercity bus and train fare, and ship fare</td>
<td></td>
</tr>
<tr>
<td>TOTHFARP</td>
<td>Tip expenditures last quarter on other transportation fares including intercity bus and train fare, and ship fare</td>
<td></td>
</tr>
<tr>
<td>TOTHRLOC</td>
<td>Total trip expenditures on lodging this quarter including rent for vacation home, and motels</td>
<td></td>
</tr>
<tr>
<td>TOTHRLOP</td>
<td>Total trip expenditures on lodging last quarter including rent for vacation home, and motels</td>
<td></td>
</tr>
<tr>
<td>TOTHTREC</td>
<td>Trip expenditures this quarter for other transportation expenses including parking fees, and tolls</td>
<td></td>
</tr>
<tr>
<td>TOTHTREP</td>
<td>Trip expenditures last quarter for other transportation expenses including parking fees, and tolls</td>
<td></td>
</tr>
<tr>
<td>TOTHVHRC</td>
<td>Trip expenditures on other vehicle rentals this quarter</td>
<td></td>
</tr>
<tr>
<td>TOTHVHRP</td>
<td>Trip expenditures on other vehicle rentals last quarter</td>
<td></td>
</tr>
<tr>
<td>TOTXEST</td>
<td>Estimated total taxes paid (FFTAXOWE + FSTAXOWE + MISCTAXX)</td>
<td></td>
</tr>
<tr>
<td>TRANSCQ</td>
<td>Transportation this quarter (CARTKNCQ + CARTKUCQ + OTHVEHCQ + GASMOCCQ + VEHFINCQ + MAINRPCQ + VEHINSCQ + VRNTLOCQ + PUBTRACQ)</td>
<td></td>
</tr>
<tr>
<td>TRANSPQ</td>
<td>Transportation last quarter (same composition as TRANSCQ)</td>
<td></td>
</tr>
<tr>
<td>TRNOTHCQ</td>
<td>Local public transportation, excluding on trips this quarter</td>
<td></td>
</tr>
<tr>
<td>TRNOTHPQ</td>
<td>Local public transportation, excluding on trips last quarter</td>
<td></td>
</tr>
<tr>
<td>TRNTRRPCQ</td>
<td>Public and other transportation on trips this quarter</td>
<td></td>
</tr>
<tr>
<td>Variable</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>TRNTRPPQ</td>
<td>Public and other transportation on trips last quarter</td>
<td></td>
</tr>
<tr>
<td>TTOTALC</td>
<td>Total of all trip expenditures this quarter (TFOODTOC + TALCBEVC + TOTHRLOC + TTRANPRC + TENTRMNPC)</td>
<td></td>
</tr>
<tr>
<td>TTOTALP</td>
<td>Total of all trip expenditures last quarter (same composition as TTOTALC)</td>
<td></td>
</tr>
<tr>
<td>TTRANPRC</td>
<td>Total trip expenditures on transportation this quarter including airfare, local transportation, tolls and parking fees, and car rentals (TGASMOTC + TVRENTLC + TTRNTRIC)</td>
<td></td>
</tr>
<tr>
<td>TTRANPRP</td>
<td>Total trip expenditures on transportation last quarter including airfare, local transportation, tolls and parking fees, and car rentals (same composition as TTRANPRC)</td>
<td></td>
</tr>
<tr>
<td>TTRNTRIC</td>
<td>Trip expenditures this quarter for public transportation, including airfares (TFAREC + TLOCALTC)</td>
<td></td>
</tr>
<tr>
<td>TTRNTRIP</td>
<td>Trip expenditures last quarter for public transportation, including airfares (same composition as TTRNTRIC)</td>
<td></td>
</tr>
<tr>
<td>TVRDIOCQ</td>
<td>Televisions, radios, and sound equipment this quarter</td>
<td></td>
</tr>
<tr>
<td>TVRDIOPQ</td>
<td>Televisions, radios, and sound equipment last quarter</td>
<td></td>
</tr>
<tr>
<td>TVRENTLC</td>
<td>Trip expenditures on vehicle rentals and other fees this quarter (TCARTRKC + TOTHVHRC + TOTHTREC)</td>
<td></td>
</tr>
<tr>
<td>TVRENTLP</td>
<td>Trip expenditures on vehicle rentals and other fees last quarter (same composition as TVRENTLC)</td>
<td></td>
</tr>
<tr>
<td>UNISTRQ</td>
<td>How many housing units, both occupied and vacant, are in this structure? (1=only other units, 2=mobile home or trailer, 3=one, detached, 4=one, attached, 5=2, 6=3–4, 7=5–9, 8=10–19, 9=20–49, 10=50 or more)</td>
<td></td>
</tr>
<tr>
<td>UTILCQ</td>
<td>Utilities, fuels and public services this quarter (NTLGASCQ + ELCTRCCQ + ALLFULCQ + TELEPHCQ + WATRPSCQ)</td>
<td></td>
</tr>
<tr>
<td>UTILLOWNC</td>
<td>Expenditures on owned vacation home utilities this quarter including water, trash, electricity, and fuels (VFUELOIC + VOTHRFC + VELECTRC + VNATLGAC + VWATERPC)</td>
<td></td>
</tr>
<tr>
<td>UTILOWNP</td>
<td>Expenditures on owned vacation home utilities last quarter including water, trash, electricity, and fuels (same composition as UTILOWNC)</td>
<td></td>
</tr>
<tr>
<td>UTILPQ</td>
<td>Utilities, fuels and public services last quarter</td>
<td></td>
</tr>
</tbody>
</table>

Wei-Yin Loh 347 GUIDE manual
<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>UTILRNTC</td>
<td>Expenditures on rented vacation home utilities this quarter including water, trash, electricity, and fuels (RFU-ELOIC + ROTHFLC + RELECTRC + RNATLGAC + R WATERPC)</td>
</tr>
<tr>
<td>UTILRNTP</td>
<td>Expenditures on rented vacation home utilities last quarter including water, trash, electricity, and fuels (same composition as UTILRNTC)</td>
</tr>
<tr>
<td>VEHFINCQ</td>
<td>Vehicle finance charges this quarter</td>
</tr>
<tr>
<td>VEHFINPQ</td>
<td>Vehicle finance charges last quarter</td>
</tr>
<tr>
<td>VEHICTAX</td>
<td>Personal property taxes for vehicles</td>
</tr>
<tr>
<td>VEHINSCQ</td>
<td>Vehicle insurance this quarter</td>
</tr>
<tr>
<td>VEHINSPQ</td>
<td>Vehicle insurance last quarter</td>
</tr>
<tr>
<td>VEHQ</td>
<td>Total number of owned vehicles</td>
</tr>
<tr>
<td>VEHQL</td>
<td>Total number of leased autos, trucks and vans</td>
</tr>
<tr>
<td>VELECTRC</td>
<td>Expenditures on electricity for owned vacation homes this quarter</td>
</tr>
<tr>
<td>VELECTRCP</td>
<td>Expenditures on electricity for owned vacation homes last quarter</td>
</tr>
<tr>
<td>VFUELOIC</td>
<td>Expenditures on fuel oil for owned vacation homes this quarter</td>
</tr>
<tr>
<td>VFUELOIP</td>
<td>Expenditures on fuel oil for owned vacation homes last quarter</td>
</tr>
<tr>
<td>VNATLGAC</td>
<td>Expenditures on natural gas for owned vacation homes this quarter</td>
</tr>
<tr>
<td>VNATLGAP</td>
<td>Expenditures on natural gas for owned vacation homes last quarter</td>
</tr>
<tr>
<td>VOTHRFLC</td>
<td>Expenditures on other fuels for owned vacation homes this quarter</td>
</tr>
<tr>
<td>VOTHRFLP</td>
<td>Expenditures on other fuels for owned vacation homes last quarter</td>
</tr>
<tr>
<td>VOTHRLOP</td>
<td>Expenditures on owned vacation homes last quarter including mortgage interest, insurance, taxes, and maintenance</td>
</tr>
<tr>
<td>VOTHRLOC</td>
<td>Expenditures on owned vacation homes this quarter including mortgage interest, insurance, taxes, and maintenance</td>
</tr>
<tr>
<td>VRNTLOCQ</td>
<td>Vehicle rental, leases, licenses, and other charges this quarter</td>
</tr>
</tbody>
</table>

Wei-Yin Loh 348 GUIDE manual
REFERENCES

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>VRNTLOPQ</td>
<td>Vehicle rental, leases, licenses, and other charges last quarter</td>
<td></td>
</tr>
<tr>
<td>VWATERPC</td>
<td>Expenditures on water and public services for owned vacation homes this quarter</td>
<td></td>
</tr>
<tr>
<td>VWATERPP</td>
<td>Expenditures on water and public services for owned vacation homes last quarter</td>
<td></td>
</tr>
<tr>
<td>WATRPSCQ</td>
<td>Water and other public services this quarter</td>
<td></td>
</tr>
<tr>
<td>WATRPSPQ</td>
<td>Water and other public services last quarter</td>
<td></td>
</tr>
<tr>
<td>WELFAREX</td>
<td>Amount received from public assistance or welfare including money received from job training grants</td>
<td>0.99</td>
</tr>
<tr>
<td>WELFREBX</td>
<td>Median of bracket range of WELFAREX</td>
<td>0.99</td>
</tr>
<tr>
<td>WHLFYR</td>
<td>Did you own any whole life insurance or other life insurance policies that can be surrendered for cash or borrowed against prior to the death of the person insured one year ago today? (1=yes, 2=no)</td>
<td>&gt;0.99</td>
</tr>
<tr>
<td>WHLFYRB</td>
<td>Range which best reflects total surrender value of these policies one year ago today (1=0–499, 2=500–999, 3=1–2.5K, 4=2.5–10K, 5=10–35K, 6=35K and over)</td>
<td>0.99</td>
</tr>
<tr>
<td>WHLFYRBX</td>
<td>Median value of bracket range for WHLFYRB</td>
<td>0.99</td>
</tr>
<tr>
<td>WHLFYRX</td>
<td>Total surrender value of these policies one year ago today</td>
<td>0.98</td>
</tr>
<tr>
<td>WHOLIFB</td>
<td>Range which best reflects the total surrender value of these policies (same codes as WHLFYRB)</td>
<td>&gt;0.99</td>
</tr>
<tr>
<td>WHOLIFBX</td>
<td>Median value of bracket range for WHOLIFB</td>
<td>&gt;0.99</td>
</tr>
<tr>
<td>WHOLIFX</td>
<td>Total surrender value of these policies as of today</td>
<td>0.98</td>
</tr>
</tbody>
</table>

References


<table>
<thead>
<tr>
<th>Code</th>
<th>PSU Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>S11A</td>
<td>Boston-Cambridge-Newton, MA-NH</td>
</tr>
<tr>
<td>S12A</td>
<td>New York-Newark-Jersey City, NY-NJ-PA</td>
</tr>
<tr>
<td>S12B</td>
<td>Philadelphia-Camden-Wilmington, PA-NJ-DE-MD</td>
</tr>
<tr>
<td>S23A</td>
<td>Chicago-Naperville-Elgin, IL-IN-WI</td>
</tr>
<tr>
<td>S23B</td>
<td>Detroit-Warren-Dearborn, MI</td>
</tr>
<tr>
<td>S24A</td>
<td>Minneapolis-St. Paul-Bloomington, MN-WI</td>
</tr>
<tr>
<td>S24B</td>
<td>St. Louis, MO-IL</td>
</tr>
<tr>
<td>S35A</td>
<td>Washington-Arlington-Alexandria, DC-VA-MD-WV</td>
</tr>
<tr>
<td>S35B</td>
<td>Miami-Fort Lauderdale-West Palm Beach, FL</td>
</tr>
<tr>
<td>S35C</td>
<td>Atlanta-Sandy Springs-Roswell, GA</td>
</tr>
<tr>
<td>S35D</td>
<td>Tampa-St. Petersburg-Clearwater, FL</td>
</tr>
<tr>
<td>S35E</td>
<td>Baltimore-Columbia-Towson, MD</td>
</tr>
<tr>
<td>S37A</td>
<td>Dallas-Fort Worth-Arlington, TX</td>
</tr>
<tr>
<td>S37B</td>
<td>Houston-The Woodlands-Sugar Land, TX</td>
</tr>
<tr>
<td>S48A</td>
<td>Phoenix-Mesa-Scottsdale, AZ</td>
</tr>
<tr>
<td>S48B</td>
<td>Denver-Aurora-Lakewood, CO</td>
</tr>
<tr>
<td>S49A</td>
<td>Los Angeles-Long Beach-Anaheim, CA</td>
</tr>
<tr>
<td>S49B</td>
<td>San Francisco-Oakland-Hayward, CA</td>
</tr>
<tr>
<td>S49C</td>
<td>Riverside-San Bernardino-Ontario, CA</td>
</tr>
<tr>
<td>S49D</td>
<td>Seattle-Tacoma-Bellevue, WA</td>
</tr>
<tr>
<td>S49E</td>
<td>San Diego-Carlsbad, CA</td>
</tr>
<tr>
<td>S49F</td>
<td>Honolulu, HI</td>
</tr>
<tr>
<td>S49G</td>
<td>Anchorage, AK</td>
</tr>
</tbody>
</table>


Wei-Yin Loh 353  GUIDE manual


REFERENCES


Tian, L., Zhao, L., and Wei, L. J. (2014). Predicting the restricted mean event time with the subject’s baseline covariates in survival analysis. Biostatistics, 15:222–233.


Wei-Yin Loh 355 GUIDE manual