Statistics 572
Statistical Methods for Bioscience II

Cecile Ane
Department of Statistics
University of Wisconsin–Madison

Spring 2010

Outline

1. Introduction
   - Course Information
   - Overview of Linear models
   - Expectations
   - Short review of simple linear regression

Outline

1. Introduction
   - Course Information
   - Overview of Linear models
   - Expectations
   - Short review of simple linear regression

Course Information

www.stat.wisc.edu/courses/st572-ane/

- Read the entire syllabus carefully.
- Complete the survey sheet.
- Switch section?
- Late homework
- Block the dates and time for the exams **NOW**:
  Tuesday, March 2
  Tuesday, April 6-13
  Friday, May 14, 12:25-2:25pm
- No discussion this week. Possible room change for discussions.

Course Information

- Get help beyond lectures: Reading materials, course website, forum, discussion sections, office hours, etc.
- Your feedback is highly appreciated. Your evaluations are most valuable to me. Ask questions, get involved! Forum on Learn UW

Course Information: Why R? Why not Microsoft Excel?

Limitations of Microsoft Excel:

- 65K “raw” data size limit
- little data protection, little/no tracking
- XL2000 has many errors, without warning. Can get negative correlation coefficients, wrong pie charts, wrong paired t-test with missing values, does not accept categorical predictors in multiple regression, etc.
- Some bugs are fixed, new bugs are created in XL2003. Still doesn’t have distributions right. Lots of errors known over 10 years without fixes.

McCullough & Wilson (2005) On the accuracy of statistical procedures in Microsoft Excel 2003. Computational Statistics & Data Analysis 49(4):1244-1252

Foresight: The International Journal of Applied Forecasting, issue 3 (2006)
R. Hesse. Incorrect Nonlinear Trend Curves in Excel
B. McCullough. The Unreliability of Excel’s Statistical Procedures
P. Fields. On the Use and Abuse of Microsoft Excel

Overview of Linear models

The course will cover: multiple regression, multi-way ANOVA (ANOVA with multiple factors), linear models with random and mixed effects, and standard experimental designs. All these are examples of linear models.

- They can give insight to biological understanding,
- observations are treated as realization from a model,
- *"All models are wrong, some are useful"* (George Box). In other words: No model accounts for all aspects of the underlying biology, but an appropriately selected model can be very useful.

Overview of Linear models

- Typically, we want to know how a response variable is related to one or more explanatory variables.
- Quantitative variables (discrete or on a continuous scale) or Categorical variables (counts in each category)
- A linear combination of the variables $X_1, \ldots, X_k$ takes the form

$$\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

- Linear and generalized linear models use linear combinations of explanatory variable to explain the response variable.

Examples of Linear models

- Linear regression

example: soybean yield, explained by hours of daylight and amount of nitrogen
response: quantitative variable Y
explanatory: one or more quantitative variable(s) $X_1, \ldots, X_k$
model: $y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + e_i$
error distribution: Normal, $e_i \sim \mathcal{N}(0, \sigma^2)$

- ANOVA

example: nitrogen level in manure, explained by diet treatment, period, and interaction
response: quantitative variable
explanatory: one or more categorical variable(s)
model: $y_i = \alpha_{j(i)} + \beta_{k(i)} + (\alpha\beta)_{j(i),k(i)} + e_i$ for instance
error distribution: Normal, $e_i \sim \mathcal{N}(0, \sigma^2)$

Examples of Linear models

- Linear regression with both types

example: milk yield, explained by diet (4 treatments) and days
response: quantitative variable
explanatory: both quantitative and categorical variables
model: $y_i = \beta_0 + \alpha_{j(i)} + \beta_1 X_{i1} + e_i$
error distribution: Normal, $e_i \sim \mathcal{N}(0, \sigma^2)$

- Polynomial regression

example: # bacterial colonies (in log CFU), explained by temperature
response: quantitative variable Y
explanatory: one quantitative variable
model: $y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_k X_i^k + e_i$
error distribution: Normal, $e_i \sim \mathcal{N}(0, \sigma^2)$

Examples of Linear models

- Mixed models

example: percentage cover of vegetation, explained by site –modeled as a random effect– and soil moisture
response: quantitative variable
explanatory: variables with both it fixed and *random* effects
model: $y_i = \beta_0 + a_{j(i)} + \beta_1 X_i + e_i$ and $a_j \sim \mathcal{N}(0, \sigma_a^2)$
error distribution: Normal, $e_i \sim \mathcal{N}(0, \sigma^2)$

- Repeated measures

example: hormone concentration, explained by individual and day
response: quantitative variable
explanatory: one or more variable, including random effect for individual
error distribution: Normal

Examples of Linear models

- Logistic regression

example: seed germination, explained by temperature and treatment
response: categorical variable with 2 levels
explanatory: one or more
model: $\mathbb{P}\{y_i = 1\}$ is a function of $\beta_0 + a_{j(i)} + \beta_1 X_{i1} + \beta_2 X_{i2}$.
error distribution: Binomial

- Poisson regression

example: number of seeds produced, explained by treatment, light intensity and age of the plant
response: discrete variable, non-negative integer valued.
explanatory: one or more variable(s)
model: $\mathbb{P}\{y_i = k\}$ is a function of $\beta_0 + a_{j(i)} + \beta_1 X_{i1} + \beta_2 X_{i2}$.
error distribution: Poisson

Data request

- I will illustrate each type of models with an example.
- Case studies will be most interesting if examples relate to your own research.
- If you, or someone in your lab has data that falls into the scope of these models and you are willing and able to share, please contact me.

Expectation –Computing: using R.

I will assume basic skills. The course webpage has resources.

R is extended with many packages developed by many people.
We will use these, and possibly others:
`lattice` has functions for graphics,
`lme4` and `nlme` have functions for mixed effects models.

It is easy to install a package, if the computer is connected to the internet: Start R. At R's command line, type:
`install.packages("lattice")` –only once.
At each session when you need the package, type:
`library(lattice)`

Expectation –Computing: using R.

Good practice: keep assignments/projects in separate folders. Keep a plain text file (`.r` extension) with the list of commands to replicate what you have done. Example...

Being able to use a computing software is essential for you to analyze your own data when the time comes.

I will expect that you will experiment with R, try things on your own, so as to get a good understanding of how R works. Getting error and warning messages is normal while experimenting.

Don’t get stuck: get help! Forum, friends, TA, instructor.

Expectation –Assignments.

Must be written clearly. When including R commands and output, don’t put them alone. Add comments to explain in English what the commands are doing, and interpret the results.

When using graphs, include axis labels, legend if necessary, etc.

Your second take-home midterm exam should look like a well-written report, that a colleague in the field should be able to understand.

FEV data set

654 observations with 5 variables. `fev`: forced expiratory volume (liters). Larger values associated with better lung function. `ht`: height (in), `age`: in years, `sex` and `smoke`.

```
fev = read.table("fev.txt", header=T)
str(fev)
with(fev, mean(age))
with(fev, sapply(split(ht,sex) , mean))

library(lattice)
xyplot(ht~age|sex, auto.key=T,group=sex,data=fev)
xyplot(fev~age|smoke, auto.key=T,group=sex,data=fev)
bwplot(fev~sex|smoke,data=fev)
bwplot(fev~smoke|age,data=fev)
bwplot(fev~smoke|equal.count(ht),data=fev)
mymat = matrix(c(3,9.5,9.5,11.5,11.5,13.5,13.5,15.5,15.5,20),
               5,2,byrow=T)
shingle(fev$age, intervals=mymat)
bwplot(fev~smoke|shingle(age, intervals=mymat),data=fev)
xyplot(fev~smoke|shingle(age, intervals=mymat),data=fev,group=sex,
       auto.key=T)
```

Simple linear regression

Let's fit a line to explain fev by height:

```
fit1 = lm(fev~ht, data=fev)
```

How do we test the presence of a trend?

```
summary(fit1)
```

Assumptions:

How do we check these assumptions?

Simple linear regression

```
layout(matrix(1:4,1,4))
plot(fit1)
```

- plot residuals versus predicted values: check that there is no curved pattern.
- qq-plot of residuals. How important is this?
- plot residuals versus predicted values: check that their variability stays the same (width around the horizontal axis)

```
plot(fev~ht, data=fev)
```