

Breiman Inaugural Award Lectures

JSM, Baltimore, Aug 1, 2017

Leo Breiman, a Giant in Statistics: Remarks

followed by “Backward multiple imputation estimation of the conditional lifetime expectancy function, with application to censored human longevity”, Kong, Klein, Klein, Lee, Wahba.

Grace Wahba

Links to these slides in my website

<http://www.stat.wisc.edu/~wahba/> – > TALKS

Leo Breiman: a Giant in Statistics

Leo Breiman was a warm hearted person happy to discuss many ideas, and there was much sadness in the Statistics community when he passed away prematurely in 2005.

To quote from his Berkeley obituary: “He ... once said he’d advise a student considering a career in statistics today to ‘remember that the great adventure of statistics is in gathering and using data to solve interesting and important real world problems ’ ”.

Leo Breiman: a Giant in Statistics (continued)

He developed Random Forests, well known to be an effective classification program, and was well known for his work on Classification and Regression Trees among other things.

These works do not have models, and in a 2001 paper in *Statistical Science* (“Statistical Modeling: The Two Cultures”) he disdained linear and related parametric models, but he went on to praise what he called “algorithmic models”. Since he said some nice things about me: “Jerome Friedman and Grace Wahba have done pioneering work on the development of algorithmic methods”, I assume he meant what are now called nonparametric, or flexible models.

Leo Breiman: a Giant in Statistics (continued)

Accordingly I've chosen in this talk to discuss some work involving nonparametric models that are motivated by particular real world problems that have come up in the Beaver Dam Eye Study.

Outline

1. The Beaver Dam Eye Study and a 24 Year Collaboration.
2. “Backward multiple imputation estimation of the conditional lifetime expectancy function with application to censored human longevity data”, Kong, Klein, Klein and Wahba, PNAS 2015.
3. Sketch of selected other results.

The Beaver Dam Eye Study and a 24 Year Collaboration

The Beaver Dam Eye Study (BDES) is an ongoing population-based study of age-related ocular disorders with baseline and follow-ups every five years. The PI's are Drs. Ron and Barbara E. K. Klein of the Ophthalmology Department at Madison. Subjects at baseline, examined between 1988 and 1990, were a group of 4,926 people aged 43–86 from Beaver Dam, Wisconsin, and a host of variables were recorded. The collaboration resulted in 12 model approaches designed to answer specific scientific questions raised by Ron and Barbara, and used to analyze the BDES data. As an example, the survival status, including ages at death, for members of this population were updated to 31 Dec 2013 with 2,014 individuals who were alive. BDES provided us an excellent opportunity to study the conditional lifetime expectancy with a model proposed by Jing Kong and collaborators.

References

- [1] G. Wahba, Y. Wang, C. Gu, R. Klein, and B. Klein. Structured machine learning for ‘soft’ classification with smoothing spline ANOVA and stacked tuning, testing and evaluation. In J. Cowan, G. Tesauero, and J. Alspector, editors, *Advances in Neural Information Processing Systems 6*, pages 415–422. Morgan Kauffman, 1994.
- [2] G. Wahba, Y. Wang, C. Gu, R. Klein, and B. Klein. Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Ann. Statist.*, 23:1865–1895, 1995. Neyman Lecture.
- [3] Y. Wang, G. Wahba, C. Gu, R. Klein, and B. Klein. Using smoothing spline ANOVA to examine the relation of risk factors to the incidence and progression of diabetic retinopathy. *Statistics in Medicine*, 16:1357–1376, 1997.

- [4] G. Wahba, X. Lin, F. Gao, D. Xiang, R. Klein, and B. Klein. The bias-variance tradeoff and the randomized GACV. In M. Kearns, S. Solla, and D. Cohn, editors, *Advances in Information Processing Systems 11*, pages 620–626. MIT Press, 1999.
- [5] X. Lin, G. Wahba, D. Xiang, F. Gao, R. Klein, and B. Klein. Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV. *Ann. Statist.*, 28:1570–1600, 2000.
- [6] F. Gao, G. Wahba, R. Klein, and B. Klein. Smoothing spline ANOVA for multivariate Bernoulli observations, with applications to ophthalmology data, with discussion. *J. Amer. Statist. Assoc.*, 96:127–160, 2001.
- [7] H. Zhang, G. Wahba, Y. Lin, M. Voelker, M. Ferris, R. Klein, and B. Klein. Variable selection via basis pursuit for non-Gaussian data. Technical Report 1042, Statistics Department University of Wisconsin, Madison WI, 2001. In Proceedings of the ASA Joint Statistical

Meetings 2001 (CDROM), available from the American Statistical Association.

- [8] H. Zhang, G. Wahba, Y. Lin, M. Voelker, M. Ferris, R. Klein, and B. Klein. Variable selection and model building via likelihood basis pursuit. *J. Amer. Statist. Assoc.*, 99:659–672, 2004.
- [9] W. Shi, G. Wahba, S. Wright, K. Lee, B. Klein, and R. Klein. LASSO Pattern search algorithm with applications to ophthalmology and genomic data. *Statistics and Its Interface*, 1:137–153, 2008. SII-1-1-A12-Shi.pdf, PMID:PMC2566544.
- [10] H. C. Bravo, K. Lee, B. E. K. Klein, R. Klein, S. Iyengar, and G. Wahba. Examining the relative influence of familial, genetic, and environmental covariate information in flexible risk models. *Proceedings of the National Academy of Sciences*, 106(20):8128–8133, 2009.
- [11] J. Kong, B. Klein, R. Klein, K. Lee, and G. Wahba. Using distance correlation and Smoothing Spline ANOVA to assess associations of familial relationships, lifestyle factors, diseases and mortality. *PNAS*, pages 20353–20357, 2012. PMID: 3528609.

- [12] J. Kong, B. Klein, R. Klein, and G. Wahba. Backward multiple imputation estimation of the conditional lifetime expectancy function with application to censored human longevity data. *PNAS*, 112:12069–12074, 2015. PMID: PMC4593088.

The Conditional Lifetime Expectancy Function

J. Kong, B. Klein, R. Klein, and G. Wahba. Backward multiple imputation estimation of the conditional lifetime expectancy function with application to censored human longevity. *PNAS*, 112:12069–12074, 2015.

Conditional lifetime expectancy is your expected lifetime, conditional on you having reached a particular age, as a function of that age, and in this paper, multiple other covariates.

The paper uses a Smoothing Spline ANOVA (SS-ANOVA) model for your achieved age and multiple covariates. The SS-ANOVA class of models was developed partly in references [1,2], 1994-5. The paper provides an improved method for multiple imputation of lifetimes for right censored subjects. An SS-ANOVA model is a function of several, say d , variables, of the form $f(t) = f(t_1, \dots, t_d)$

$$f(t) = \mu + \sum_{\alpha} f_{\alpha}(t_{\alpha}) + \sum_{\alpha < \beta} f_{\alpha\beta}(t_{\alpha}, t_{\beta}) + \sum_{\alpha < \beta < \gamma} f_{\alpha\beta\gamma}(t_{\alpha}, t_{\beta}, t_{\gamma}) + \dots \quad (1)$$

where the elements in the expansion are made unique and in practice, the expansion is truncated in some manner. Components which are continuous variables are often represented by splines.

Our estimate of the conditional lifetime expectancy function is based on 4,926 people in BDES at baseline and their baseline covariates, and their ages of death updated to 31 December 2013. 2014 people were still alive, so their age at death is right censored.

TABLE 1: *Variable description in the BDES SS-ANOVA model*

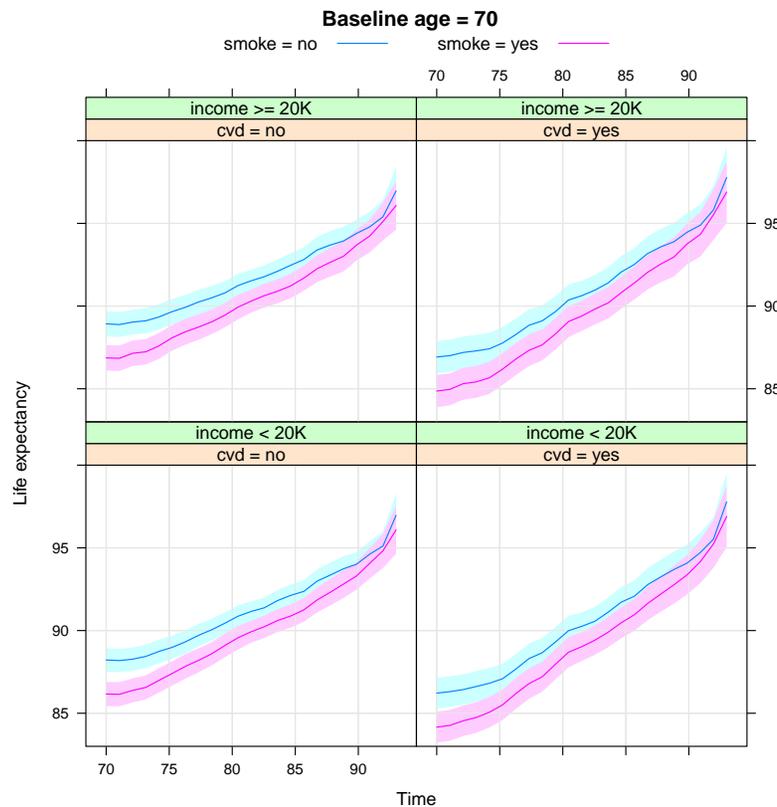
variable	units	description
lastage	years	censored age at death
survflag	yes/no	survival indicator
baseage	years	age at baseline
gender	F/M	gender
edu	years	highest year school/college completed
bmi	kg/m ²	body mass index
smoke	yes/no	history of smoking
inc	yes/no	household personal income > 20K
diabetes	yes/no	history of diabetes
cancer	yes/no	history of cancer
heart	yes/no	history of cardiovascular disease
kidney	yes/no	history of chronic kidney disease

SS-ANOVA Model for (imputed) lastage (Deathage).

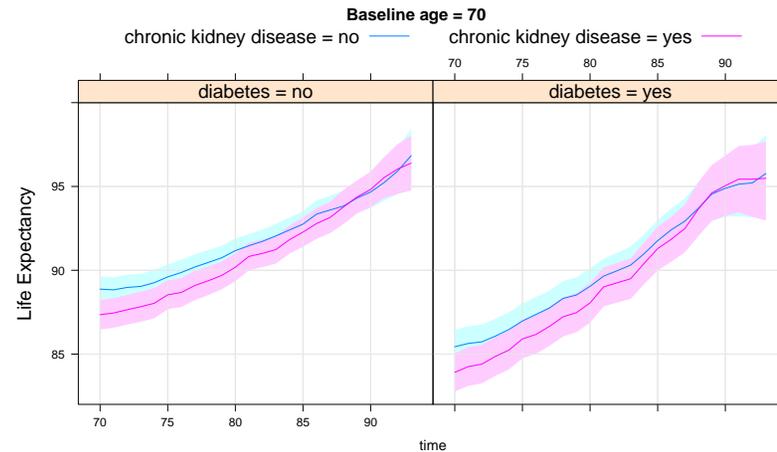
$$\begin{aligned}(\textit{imputed}) \textit{lastage} = & \mu + f_1(\textit{baseage}) + \beta_{\textit{gender}}I_{\{\textit{gender}=F\}} \\ & f_2(\textit{edu}) + f_{12}(\textit{baseage} : \textit{edu}) + \\ & f_3(\textit{bmi}) + \beta_{\textit{smoke}}I_{\{\textit{smoke}=no\}} \\ & \beta_{\textit{inc}}I_{\{\textit{inc}>20K\}} + \beta_{\textit{diabetes}}I_{\{\textit{diabetes}=no\}} + \\ & \beta_{\textit{cancer}}I_{\{\textit{cancer}=no\}} + \beta_{\textit{heart}}I_{\{\textit{heart}=no\}} + \\ & \beta_{\textit{kidney}}I_{\{\textit{kidney}=no\}}\end{aligned}$$

Functions f_1, f_2 and f_3 are cubic smoothing splines and f_{12} is the tensor product of two cubic smoothing splines. The remaining covariates are unpenalized and modeled as linear terms with $I_{\{.\}}$ as indicator functions.

An imputation scheme is presented, which makes use of the covariates to obtain imputed data $(y_i, x_i) | y_i > t$. From the SS-ANOVA Model fits for the expected lifetime given the covariates, and using imputed data one obtains the conditional lifetime expectancy function.

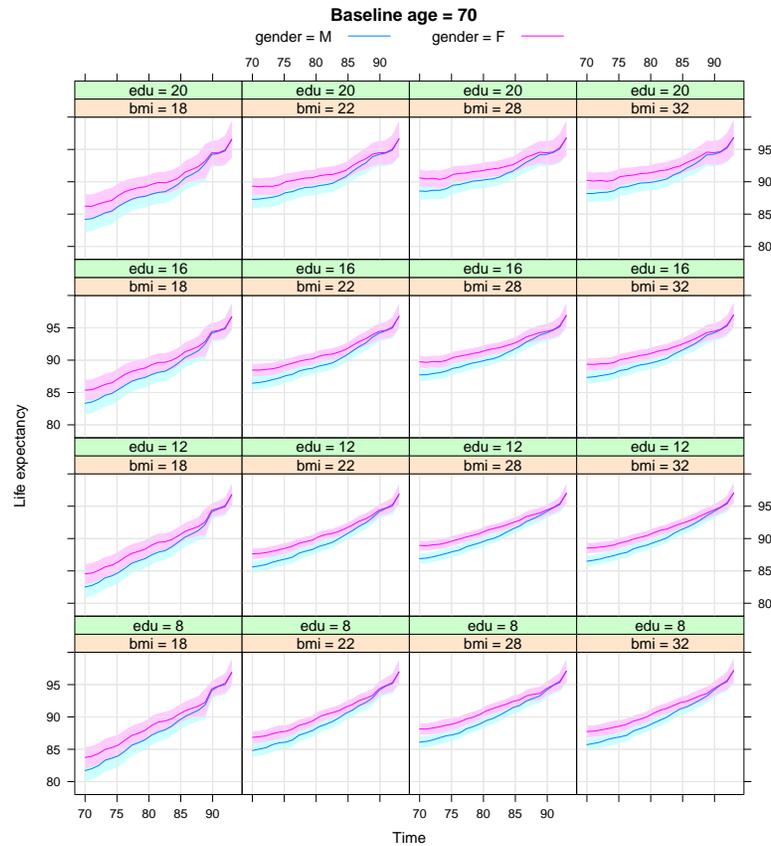


Conditional lifetime expectancy function for women with baseage=70, bmi=28, edu=12, no disease other than heart (cvd). By smoking (blue=NO, pink=YES), heart (left col=NO), income (top row >20K). The x-axis goes from $t = 70$ to $t = 93$. The y-axis is $\hat{e}(t|X = x)$smoking-bad, cvd-bad, higher income-somewhat good...



Conditional lifetime expectancy function for women with baseage=70, smoking=NO, bmi=28, income > 20K, edu=12, and no heart disease or cancer. By diabetes (left=NO) and kidney disease (blue=NO, pink=YES). The x-axis goes from t = 70 to t = 93. The y-axis is $\hat{e}(t|X = x)$.

...diabetes-very bad, kidney disease-bad. Note merging at higher ages...



Conditional LEF for M, F with baseage=70, smoking=NO, income > 20K, no disease, M = blue F = pink, bmi-rows increasing l to r, edu-columns decreasing top to bottom. x and y axes same as before ..F-better, higher ed-good (top row), bmi-midlevel better (cols 2,3)..

A Few Selected Medical Results

- [5] Retinal pigmentary abnormalities in women at baseline, predictors horm, hist, bmi, age, sysbp, chol. **Hormones protective, high cholesterol protective (!)**
- [10] Retinal pigmentary abnormalities in women at baseline. Same predictors as [5] plus two snps, pedigree information and smoking. **These new predictors all add information.**
- [9] Five year myopic change in persons 60-69 at baseline. **Heavy smokers who take vitamins have a smaller risk of myopic change, across levels of other variables, while for non-smokers, taking or not taking vitamins does not change risk.**
- [11] Subjects who have died by March 2011 (n=1004). predictors baseage, gender; education, bmi, smoking, income; cancer, diabetes, heart, kidney; and pedigrees. **Mortality runs in families, as does education, bmi, smoking and income**

Summary

We have developed and tested a number of new or improved statistical methods for analyzing complex data sets with a variety of heterogeneous predictor variables, some interacting. It's been an extraordinary privilege to work on BDES data, a study that is the gold standard for well designed demographic studies.

It is an honor to be here for this session in which we remember Leo Breiman and his groundbreaking contributions to the toolkit of the statistical data analyst.