Twenty-four years of collaboration

And the latest example: "Backward multiple imputation estimation of the conditional lifetime expectancy function, with application to censored human longevity".

Klein Group Data Meeting

Grace Wahba

February 4 , 2016

Introduction

These overheads are extracted from a talk that I gave in Dr. Ron Klein and Dr. Barbara E. K. Klein's Group Data Meeting, celebrating 24 years of collaboration. The Kleins are members of the Ophthalmology Department of the School of Medicine and Public Health at the University of Wisconsin-Madison. This collaboration resulted in 12 papers in refereed journals, several of them winning accolades of various sorts. The first was published in 1994 and the most recent in September 2015. These twelve are all coauthored with Ron and Barbara, and several also with Kris Lee of their group. I owe much thanks to Ron, Barbara and Kris, and to Scot Moss, Heidi Christian, the staff of the fundus center, and others who have asked questions or made remarks over the years that have caused us to think hard and improved the work.

Outline

1. The Papers (briefly annotated)
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

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. Tesauro, 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, JSM 1994.

[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. Full Oral Presentation (24/474) NIPS 1998

[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. Wald Lectures, JSM 2003

[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, PMCID:PMC2566544. Presidents Invited Address JSM 2007

[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. Featured in "In This Issue", Ilsong Lecture, Seoul, 2008, Fifth International Congress of Chinese Mathematicians 2010

[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. PMCID: 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. *PNAS*, 112:12069–12074, 2015. Fisher Lecture JSM 2015

The Conditional Lifetime Expectancy Function

This is our latest paper:

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 coveriates. The SS-ANOVA class of models was developed partly in references [1,2], 1994-5. In today's paper an (original) 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
$f(t) = f(t_1, ..., t_d)$

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

where the elements in the expansion are made unique and 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 2function is based on 4,926 people in BDES at baseline and their baseline covariates, and their ages of death updated by 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*

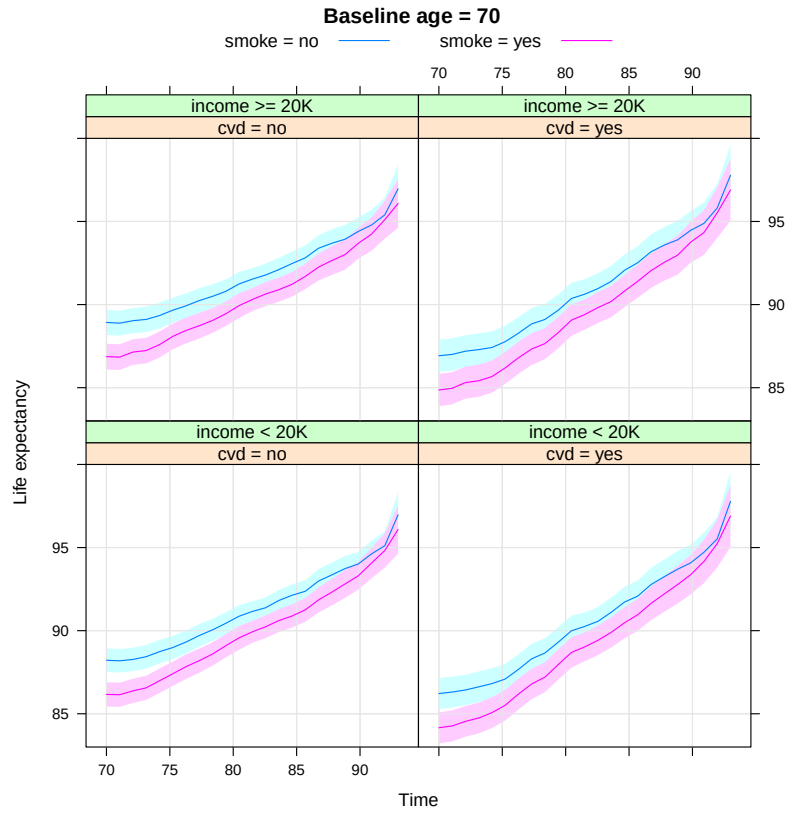
variable	units	description
lastage	years	censored age at death
survflag	yes/no	survival indicator
baseage	years	age at baseline
gender	F/M	gender
edu	years	highest year school/college completed
bmi	kg/m^2	body mass index
smoke	yes/no	history of smoking
inc	yes/no	household personal income $>$ 20K
diabetes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

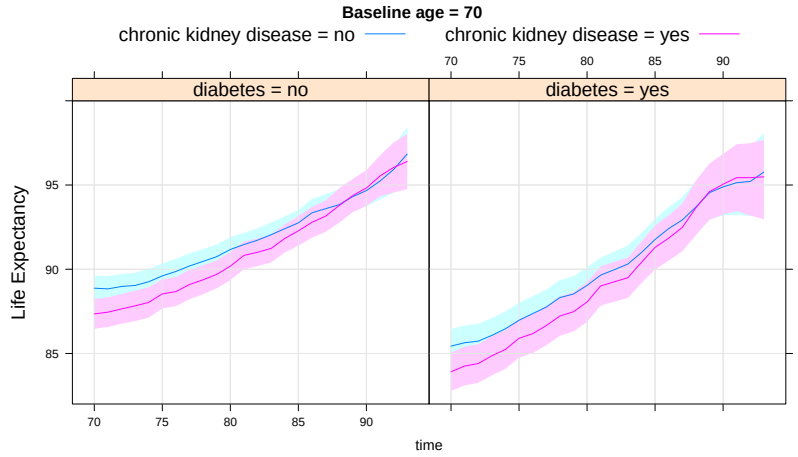
$$
\begin{aligned}
(imputed)\ lastage = \mu + f_1(baseage) + \beta_{gender} I_{\{gender=F\}} \\
& f_2(edu) + f_{12}(baseage : edu) + \\
& f_3(bmi) + \beta_{smoke} I_{\{smoke=no\}} \\
& \beta_{inc} I_{\{inc>20K\}} + \beta_{diabetes} I_{\{diabetes=no\}} + \\
& \beta_{cancer} I_{\{cancer=no\}} + \beta_{heart}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

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.

From the SS-ANOVA Model fits for the expected lifetime given the covariates, using imputed data $(y_i, x_i)|y_i > t$, one obtains the conditional lifetime expectancy function.

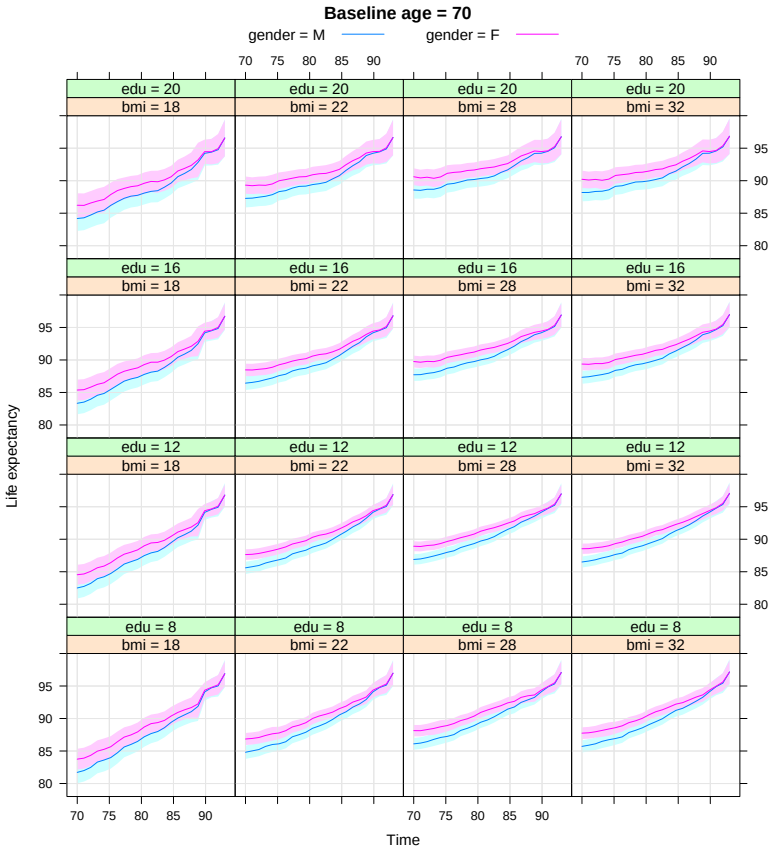


Conditional lifetime expectancy function for women with baseage=70, bmi=28, edu=12, no disease other than heart. By smoking (blue=NO), heart (left col=NO), income (top row >20K). The x-axis goes from t = 70 to t = 93. The y-axis is $\hat{e}(t|X = x)$. ...smoking-bad, history of CVD-bad, higher income-somewhat 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). 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 persons with baseage=70, smoking=NO, income $> 20K$, no disease, M = blue, 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 heterogenous variables and pedigree or other pairwise information. Its been a extraordinary privilege to work on BDES data, a study that is the gold standard for well designed demographic studies.