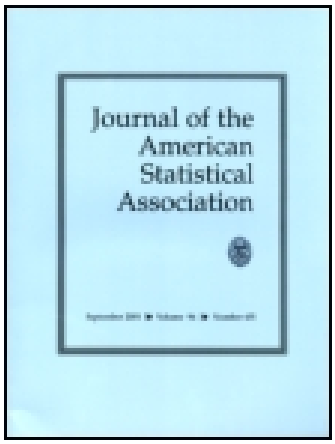

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/uasa20

Fixed Effects Regression Methods for Longitudinal Data Using SAS

John M Neuhaus[a]
[a] University of California, San Francisco
Published online: 01 Jan 2012.

To cite this article: John M Neuhaus (2006) Fixed Effects Regression Methods for Longitudinal Data Using SAS, Journal of the American Statistical Association, 101:475, 1308-1308, DOI: 10.1198/jasa.2006.s118

To link to this article: http://dx.doi.org/10.1198/jasa.2006.s118

Fixed Effects Regression Methods for Longitudinal Data Using SAS.

Paul D. ALLISON. Cary, NC: SAS Institute, 2005. ISBN 1-59047-568-2. vi + 148 pp. $34.95 (P).

Allison's objective in this book is to convince the reader that fixed-effects models and methods (models that contain fixed, subject-specific intercepts) can produce highly effective analyses of longitudinal data. In particular, Allison seeks to demonstrate that fixed-effects methods can control for all possible subject-level characteristics, measured or unmeasured, as long as they do not vary over time. This objective is not as outrageous as it may seem; it is well known that conditional likelihood methods for canonical-link generalized linear models achieve this control and, as the author illustrates in the book, fixed-effects and conditional likelihood methods are closely related.

Allison's approach in this book is similar to that in his previous SAS books on survival analysis (Allison 1995) and logistic regression (Allison 1999). He builds up theory for a particular response type in some simple settings and then illustrates results and properties of the methods using worked examples, which include detailed SAS programs and output. I found Allison's survival analysis book useful in courses for graduate students in epidemiology and related biomedical fields and believe that such students will find the current text useful as well.

The book considers the analysis of longitudinal or clustered response data in settings where scientific interest focuses on the association of time-varying, within-subject (cluster) covariates with a repeated response and the investigator wants to control for the potential confounding effects of time-invariant or baseline variables. Data analysts commonly face such problems, so the book addresses a useful area. The book treats different response types in separate chapters, because the fitting methods and their properties differ by response type.

Chapter 2 demonstrates that for identity link models for continuous responses, three distinct approaches, maximum likelihood/least squares estimation of the model that includes fixed intercept parameters for each subject (cluster), conditional likelihood estimation, and an approach that partitions the time-varying covariates into between- and within-cluster components, all yield identical estimates and standard errors of the associations of interest. More importantly, Allison points out that these three methods control for the potential confounding effects of all time-invariant or baseline variables, measured or unmeasured. Chapter 2 also points out that standard linear mixed-effects models that include random intercepts do not produce estimates identical to the those of other three methods, because this approach controls only for the confounding effects of those variables explicitly included in the model. Chapter 2 serves as a template for the subsequent chapters, which examine the performance of the aforementioned four methods: (1) models that include fixed, subject-specific intercepts; (2) conditional likelihood approaches; (3) models that decompose covariates into between- and within-cluster components; and (4) mixed-effects models that include random intercepts.

Chapter 3 considers repeated binary and categorical response data and methods based on logistic regression. Chapter 3 demonstrates that although methods (2) and (3) continue to provide accurate estimation with categorical responses, methods that include fixed, subject-specific intercepts yield highly biased estimates of the associations of interest. Allison also shows that fitting categorical response models is more complicated and time-consuming in SAS than fitting continuous-response models, which leads him to make some questionable analytic recommendations. To decrease computation time for between- or within-cluster covariate methods, Allison suggests fitting marginal models based on the generalized estimating equation (GEE) approach, although he acknowledges that in the presence of time-varying covariates, this will produce estimates of associations that are not typically of scientific interest. Decreasing computation time also motivates the suggestion to consider methods based on penalized quasi-likelihood (PQL) and the GLIMMIX routine in SAS. However, it is well known that PQL can provide highly biased parameter estimates. In addition, for most routine applications, computing times for the between- or within-cluster covariate approach are not excessive and will continue to decrease as processors improve. Therefore, computing speed is not a compelling basis on which to choose the analytic method. Chapter 3 would have been clearer and more effective had it only discussed subject-specific approaches and avoided the somewhat subtle distinction between these and population-averaged models.

Chapter 4 considers repeated-count data and methods based on Poisson and negative binomial regression. Allison demonstrates that the four analytic approaches of interest perform similar to those for continuous responses. This result is consistent with existing literature that indicates analogous performance of identity and log-link models for clustered data. In particular, for either repeated Gaussian or Poisson responses, methods (1) and (2) yield identical estimates and standard errors of the parameters of interest. However, Chapter 4 notes that count data frequently exhibit excess variation relative to the Poisson model, and Allison accommodates this overdispersion using negative binomial models. Although conditional likelihood methods do not exist for negative binomial outcomes, Chapter 4 provides empirical evidence to indicate that approaches (1) and (3) provide identical estimates and standard errors of the parameters of interest. Little theory exists to compare these two approaches in the negative binomial model context. As in Chapter 3, computation times to fit mixed-effects negative binomial models with between- and within-cluster covariate decompositions can be large, leading Allison to again suggest a GEE-based estimation strategy. The aforementioned criticisms of this strategy apply equally to count and binary data. However, it is true that subject-specific and population-averaged parameters differ less in magnitude for negative binomial data than for binary data. Regardless, one should fit models that estimate quantities of scientific interest and not consider expedient approaches that do not estimate these quantities.

Chapter 5 considers repeated time-to-event outcomes and methods based on the Cox proportional hazards model. Allison presents empirical results to suggest that the four methods perform similar to the case of repeated binary responses, reflecting the connection between logistic regression and survival analysis. However, theory and analytic results for repeated time-to-event outcomes are less developed than those for binary outcomes. Analogous to conditional likelihood methods for binary data, Chapter 5 notes that one can effectively control for all time-invariant confounders using stratified Cox models. Also, as with logistic regression, adding fixed, subject-specific intercepts to Cox models does not produce consistent estimators of parameters of interest. Survival analysis analogs of mixed-effects models are Cox models with frailties, but Chapter 5 does not mention these, perhaps due to lack of software in SAS. This precludes the author from examining the performance of survival analysis analogs of the hybrid approach that decomposes covariates into between- and within-cluster components. The chapter mentions that the author has made some attempts at implementing a hybrid approach, but it is not clear exactly what was done. The chapter also suffers from a scarcity of references to relevant theory.

Chapter 6 casts the linear mixed-effects models of Chapter 3 as structural equation models and illustrates that the latter models can accommodate data features, such as endogenous variables, that the former models cannot. Chapter 6, the book's shortest chapter, merely provides an introduction to structural equation models, but it demonstrates how to fit a few such models using SAS procedures.

Overall, I found this to be an interesting and practical book that should be of interest to analysts of clustered data. The utility and clarity of the methods and results are greatest for continuous responses (Chap. 2) and decrease as the reader moves through subsequent chapters. The book emphasizes that mixed-effects methods that decompose covariates into between- and within-cluster components and that conditional likelihood methods for canonical link generalized linear models typically yield estimates of quantities of interest that are free of the confounding effects of time-invariant variables, measured or unmeasured. This is a useful result for longitudinal data analysts to keep in mind. The suggestions to reduce computation time by estimating the effects of time-varying covariates using GEE methods are not helpful and confuse the distinction between subject-specific and population-averaged methods; quickly computing the wrong answer is not useful.

John M. NEUHAUS
University of California, San Francisco

REFERENCES

Allison, P. (1995), *Survival Analysis Using the SAS System: A Practical Guide*, Cary, NC: SAS Institute.

——— (1999), *Logistic Regression Using the SAS System: Theory and Application*, Cary, NC: SAS Institute.

Nonparametric and Semiparametric Models.

Wolfgang HÄRDLE, Marlene MÜLLER, Stefan SPERLICH, and Axel WERWATZ. New York: Springer, 2004. ISBN 3-540-20722-8. xxvii + 299 pp. $89.95.

One common problem for researchers is determining whether statistical assumptions hold. Standard courses in linear models stress plots and residual