The Stata Journal (2003)
3, Number 2, pp. 208–210

Review of Generalized Estimating Equations by Hardin and Hilbe

Steven Stillman
Labour Market Policy Group, New Zealand Department of Labour
steven.stillman@lmpg.dol.govt.nz

Abstract. The new book by Hardin and Hilbe (2003) is reviewed.

Keywords: gn0008, generalized estimating equations, generalized linear models

1 Introduction

Hardin and Hilbe (2003) have written a very detailed book on the statistical methodology of generalized estimating equations (GEE). This work is very much a continuation of their previous book (Hardin and Hilbe 2001), which focused on generalized linear models (GLM): for a review in this journal, see Newson (2001). GEE is an extension to GLM that does not require independent observations and thus can be used to analyze clustered and longitudinal data. At the simplest level, a variance–covariance matrix, which describes the correlation between observations, is specified, and multivariate weighted least squares is used to estimate a GEE model.

The book primarily focuses on explaining this process in detail. It assumes that the reader has a fundamental understanding of GLM, maximum likelihood estimation, and distributional statistics. As Hardin and Hilbe are the original authors of the `glm` and `xtgee` commands in Stata, which are used to estimate GLM and GEE models, respectively (Hilbe 1993), Stata is often used in the book to present various illustrative examples. The authors also provide detailed information on estimating these models in SAS, S-Plus, and SUDAAN.

2 Detailed comments

The book contains four main chapters, with a fifth chapter that describes various example datasets and useful programs for estimating GEE models. The first chapter provides an introduction to the subject matter. It is both detailed and technical, and it immediately becomes clear to the reader how much technical understanding of GLM and distributional statistics is expected by the authors. Unfortunately, this limits the reach of the book, as many social scientists, even those with strong empirical training, have limited experience with GLM methodology. This chapter also reviews different software packages that can estimate GLM and GEE models. Considering the authors' experience writing the main Stata programs for these models, the reviews are very neutral and even-handed.

The second chapter provides a comprehensive description of GLM estimation, but again at a level that requires strong prior knowledge of the methodology. The chapter begins by reviewing general maximum-likelihood estimation methods. Given the high level of statistical knowledge assumed by the authors in most of the book, this review seems unnecessary. While the description of GLM is comprehensive, little depth is provided, especially on how a researcher chooses between different models. For example, the authors describe the choice between random- and fixed-effects models as depending mainly on whether covariates are constant within panels. No mention is made of the fundamental difference between these models: random-effects models are inconsistent when covariates are correlated with the unobserved individual error term.

The third chapter describes GEE estimation in detail and is the core chapter of the book. As discussed above, the first step in estimating a GEE model is specifying a variance-covariance matrix that describes the correlation between observations. The beginning of the chapter is organized around the different correlation matrices that can be specified. For each type of structure, the authors demonstrate how the GEE model is estimated using both an example estimated by hand and one estimated using the `xtgee` command in Stata. The authors do a good job describing each type of correlation structure and the actual process of estimating each model. The chapter then continues by running through new developments that have extended GEE estimation to multinomial, robust, instrumental variable, and missing data models. In the last section of the chapter, the authors discuss how to choose an appropriate model.

As with the GLM chapter, I found the discussion on choosing an appropriate model to be lacking. The authors write that the choice depends on the scientific question of interest and on the characteristics of the data. While I would agree that these are important factors in model selection, in my mind, it is how the fundamental assumptions of a particular model interact with these factors that determines model choice. Importantly, the authors do not discuss these fundamental assumptions. They do not explain whether the random effects assumptions (i.e., that covariates are uncorrelated with the individual error term) are necessary for consistent estimates. In addition, no mention is made of whether choosing the wrong error structure leads to inconsistent estimates, biased standard errors, or just inefficient estimates.

The fourth chapter further discusses model choice, focusing on residual analysis and model goodness of fit. The techniques examined in this section are mainly general ones that apply to regression methods other than GEE. These include using information criteria and graphical analysis of residuals to determine best model fit and using deletion methods and influence measures to identify data outliers. Given the generality of this chapter, I did not find that it added much to the book.

3 Summary

This book is a comprehensive summary of the methodology of generalized estimating equations. The authors state that the intended audience for this book includes active researchers as well as theoretical statisticians. I am not sure how much it will appeal

to either group. My main statistical training has been in econometrics, and hence my knowledge about GEE was fairly limited before reading this book. Now I have a much better understanding of how GEE models are estimated in practice. However, I have not gained any insight into why, as a researcher, I should want to use this technique. Not enough information is given in the book on the assumptions that are required to estimate GEE models, the advantages that they have over maximum likelihood models, and what limitations are inherent in using this method. I am guessing that theoretical statisticians, on the other hand, will be frustrated by the lack of rigor in the book. While all of the important formulas are presented, the book does not delve into the theoretical underpinnings of GEE. No proofs are given, and simple examples only are used to demonstrate the mechanics of particular formulas. Having said this, I think that this book would make excellent supplemental reading material in a graduate-level quantitative methods course, as it provides a more comprehensive description of GEE than is, to my knowledge, available from any other source.

4 References

Hardin, J. W. and J. M. Hilbe. 2001. *Generalized Linear Models and Extensions*. College Station, TX: Stata Press.

—. 2003. *Generalized Estimating Equations*. Boca Raton, FL: Chapman & Hall/CRC.

Hilbe, J. 1993. sg16: Generalized linear models. *Stata Technical Bulletin* 11: 20–28. In *Stata Technical Bulletin Reprints*, vol. 2, 149–159. College Station, TX: Stata Press.

Newson, R. 2001. Review of Generalized Linear Models and Extensions. *Stata Journal* 1(1): 98–100.

About the Author

Steven Stillman is a Senior Research Economist in the Labour Market Policy Group of the New Zealand Department of Labour. He is also an affiliated Research Fellow at the Institute for the Study of Labor (IZA), the William Davidson Institute, and the Motu Economic and Public Policy Research Trust. His current research examines the effect of public policy and institutions on various dynamic aspects of household well-being in New Zealand, Russia, and the United States.