The Stata Journal (2006)
6, Number 2, pp. 273–278

Review of Regression Models for Categorical Dependent Variables Using Stata, Second Edition, by Long and Freese

Richard Williams
Department of Sociology
University of Notre Dame
Notre Dame, IN
richard.a.williams.5@nd.edu

Abstract. This article reviews *Regression Models for Categorical Dependent Variables Using Stata*, Second Edition, by Long and Freese.

Keywords: gn0032, categorical data, regression models

1 Introduction

I once heard a speaker say that "I research those things I know about and teach about those things I want to learn." With that thought but with some trepidation, I offered my first-ever class in categorical data analysis a year ago. Although I had at least some familiarity with most of the major models in the field, there were several I had never worked with myself, nor was I familiar with much of the software needed to estimate them. Nonetheless, this course proved to be one of the most popular I have ever taught and helped to inspire new research of my own. Much of the credit for the course's success goes to an outstanding book I relied heavily on: J. Scott Long and Jeremy Freese's (2003) *Regression Models for Categorical Dependent Variables Using Stata*, Revised Edition. I therefore eagerly accepted the offer to review the 2006 Second Edition.

The second edition's strategy is basically the same as the first's. After some introductory material, models for binary, ordinal, nominal, and count outcomes each have their own chapters. The book intuitively and nonmathematically introduces and explains the substantive rationale for each model. Although not intended to be thorough, these discussions generally do an excellent job of making the basic principles clear. The authors then show how, through postestimation procedures, hypotheses can be tested and results made more interpretable, using both their own programs and Stata's. (Even if you never read the book, you should at least install SPost9, which is one of the most extensive suites of Stata utilities written by anyone outside StataCorp.) Alternative approaches are at least briefly mentioned. Throughout, there is a strong practical and applied emphasis, as the authors present clear examples using Stata that can be easily replicated by readers who want to work along with the book.

Even without opening the book, however, you see that this is not simply a cosmetic update. The second edition is more than 150 pages longer than its predecessor. As noted in the preface (xxx), new to this edition are discussions of "the zero-truncated negative binomial models, the hurdle model for counts, the stereotype logistic regression model, the rank-ordered logit model, and the multinomial probit model." Other sections have been revised, and several useful commands have been added to SPost9.

2 Contents

Long and Freese's book is divided into three sections. The first section provides background material. Chapter 1 presents a brief overview of the book and shows how to access and install the SPost9 programs. Chapter 2 is the obligatory introduction to Stata. It shows how to access and manipulate data, run commands, and create graphics. Although experienced Stata users will probably want to skim over this section, beginners should find it more than satisfactory for their immediate needs.

Chapter 3 covers estimation, testing, fit, and interpretation. By this point, readers may be feeling anxious to get to the actual models, and the discussion here may be difficult to appreciate fully before seeing the more specific examples that follow later. Nonetheless, the chapter is worth careful attention because it introduces valuable material used throughout the book. This chapter explains the common features of maximum likelihood estimation commands and output, and it outlines ways of reformatting output via `estimates table` and `estout` (Jann 2005). The authors introduce several of their most frequently used commands. `listcoef` can be used to transform coefficients to aid interpretation: for example, coefficients can be standardized in various ways or expressed as multiplicative changes in the odds or expected counts. The chapter explores postestimation Wald and likelihood-ratio tests and discusses measures of fit. Particularly useful here is the presentation of many pseudo-R^2 measures that have been proposed for categorical data analysis, along with the Akaike and Bayesian information criteria. These measures can all be calculated with the `fitstat` command, which I suspect is one of the most popular utilities in the book.

The authors then present several aids to interpretation. Many of these involve using predicted values to create substantively meaningful profiles that illustrate the effects of the independent values. That is, the user can vary the values of the independent variables to see how predicted outcomes are affected. To aid in this task, the `prvalue` command (similar to Stata's `adjust` command but more powerful) computes predicted values of the outcomes for specified values of the independent variables. Also `prchange` computes several preprogrammed discrete and marginal changes in the marginal outcomes, showing, for example, the effect that a one-unit or 1–standard deviation change in an independent variable has on the probability of different outcomes occurring.

The second section is "Models for Specific Kinds of Outcomes". Chapter 4 presents models for binary outcomes. The primary emphasis is on the logit and probit models, with brief mention of other alternatives at the end. This chapter shows that the latent-variable and nonlinear probability models are conceptually different but mathematically

equivalent ways of motivating binary regression models. After presenting and explaining typical output from the logit and probit commands, the authors show how to perform Wald and likelihood-ratio chi-squared tests on individual and multiple coefficients. They then show how to examine residuals and potentially problematic cases, a task that is facilitated with graphics and the authors' `leastlikely` command. Long and Freese use commands like `fitstat`, `prvalue`, `prchange`, and `listcoef` to clarify ideas from chapter 3 that may have seemed a bit abstract. When I taught my class using the earlier version of this book, chapter 4 took longer than I expected. But once students were familiar with the various analytic strategies and commands, later chapters went much more quickly because they used similar ideas and often identical commands.

Chapter 5 discusses models for ordinal outcomes. Those who have mastered chapter 4 will find chapter 5 a straightforward extension of the ideas developed earlier, with most of the same commands and procedures being used in slightly new ways. The ordinal logit and ordinal probit models receive the most attention, although the chapter briefly covers other alternatives such as the generalized ordered logit model and the continuation-ratio model.

Up to this point, the changes from the earlier edition have been modest. Chapter 6, however, provides a much more extensive discussion of models for nominal outcomes with case-specific data. Particularly welcome are the new discussions on multinomial probit and stereotype logistic regression, which are estimated by using Stata 9's `mprobit` and `slogit` commands. However, the primary emphasis continues to be on the multinomial logit model fitted by `mlogit`. The authors introduce a major command of their own here, `mlogtest`. Although the same tests can be done with Stata's `test` and `lrtest` commands, `mlogtest` is much less tedious. It can automate the testing of individual variables, test whether alternatives can be combined, and do Hausman and Small–Hsiao tests of whether the assumptions behind the model are met. This chapter also introduces the `mlogview` and `mlogplot` commands, which provide a visual means for seeing the effects of changes in the independent variables.

Chapter 7 (whose material was combined with chapter 6 in the last edition) focuses on models for nominal outcomes with alternative-specific data, in which characteristics of the alternatives vary across cases. For example, if the outcome is the mode of transport that a person uses to get to work, the amount of time for a given mode of travel can differ for each person. Such models require datasets that are organized differently from the other models in the book, a task that is facilitated by the authors' new `case2alt` command. The authors address the conditional logit model (fitted by `clogit`), the alternative-specific multinomial probit model (fitted by `asmprobit`), and rank-ordered logistic regression model (fitted by `rologit`). My own lack of prior familiarity with these models made this the most challenging chapter in the book for me, but it was still clear and understandable.

Chapter 8, on models for count outcomes, was one of the clearest and most enjoyable. I found it so partly because I find count models easier to interpret and understand than other types of categorical data models, and partly because of the clarity of the authors' presentation. Stata 9 added new commands for these models, and the book reflects

this. The chapter begins by explaining the Poisson distribution and Poisson regression and shows how `prcounts` can be used to compute predicted values and help assess the fit of the model. The authors then discuss alternatives that can be used when the assumptions of Poisson regression are violated. Sometimes there is unobserved heterogeneity among the observations; the negative binomial regression model (fitted by `nbreg`) addresses this. Another problem is truncated counts, which occurs when observations with values equal to zero are excluded from the data; e.g., the number of hospital visits may be measured only for people who have visited the hospital at least once. Zero-truncated Poisson models (fitted by `ztp`) can be used in such cases. Hurdle models and zero-inflated count models deal with still other problems. As usual, the authors discuss each of these and present ways for testing the models and interpreting their results. Particularly useful is a command new to this edition, `countfit`, which explicitly compares the different models. (Researchers may also be interested in the several hurdle programs available from SSC that were written by Joseph Hilbe after the book was released.)

Chapter 9 addresses a variety of new topics. Much of the discussion focuses on how to fit, interpret, and test models with categorical independent variables. The chapter briefly mentions interaction effects and nonlinear models (e.g., categorical models with independent variables like age-squared). The authors close with a few guidelines for those who want to modify the SPost routines and suggestions for using Stata more efficiently.

The last section of the book is the appendices, which provide the syntax for the SPost commands and describe the datasets used in the examples. Many will find this section a useful reference long after having finished reading the rest of the book.

3 Strengths and weaknesses

Long and Freese stress early on that their book should not be the sole source of information for the models they cover. Long's (1997) similarly organized but far more mathematical *Regression Models for Categorical and Limited Dependent Variables* is one candidate for a companion text. However, if researchers could read only one book, many would judge Long and Freese's the most useful. I found the book's intuitive and applied approach much easier to understand and follow than those books that took a more mathematical approach.

Even more invaluable are the book's aids to interpretation. All too often, when reviewing articles, I criticize authors for discussing the signs and statistical significance of their coefficients while providing little feel for the coefficients' substantive significance. By showing how changes in independent variables affect the probabilities of events occurring, the SPost routines provide a much more intuitive way of assessing and understanding variable effects.

Overall, these strengths of the book greatly outweigh its weaknesses. Nonetheless, there were areas where I wish the authors had said a little more. Although they define

the various pseudo-R^2 and information measures produced by `fitstat`, they give little guidance in choosing between them. Also, while the authors cannot cover every possible model, it would have been helpful if in the brief discussions of other alternatives they had sometimes said more about why and when we would want to consider them. The `intreg` command offers an attractive alternative for modeling certain types of ordinal dependent variables, and I hope that a future edition of the book addresses it. Allison (1999) has warned that interaction effects in logit and probit models can be misleading, but Long and Freese never mention this issue or what to do about it.

Another weakness, more with the programs than with the book, is that user-written routines do not usually work with SPost9. The authors readily admit the problem but properly point out that their time is limited and that it is difficult for them to support commands they have not worked with themselves. Nonetheless, limited generic support by commands like `fitstat` and `prvalue` may be feasible, accompanied with appropriate "use at your own risk" warnings. Programmers might also want to consider whether they can trick SPost9 into thinking that a supported command has been used. For example, my own `gologit2` command (Williams 2006) is not currently supported by SPost, but it was relatively easy to convince SPost that the estimates were generated by `gologit` (Fu 1998). There are several other items that could be added to a programming wish list, but rather than impose these demands on the authors, StataCorp may wish to consider adding some of the SPost utilities to official Stata.

4 Conclusion

As Schumm (2005, 599) recently pointed out, "Too often, courses in applied statistics and data analysis are taught using a text covering the theory behind the methods but expecting students to pick up the details of a particular software package on their own via hastily assembled or otherwise inadequate resources." This trend will certainly not be true of any course taught using Long and Freese's book. Further, this is not just a book that shows how to use a statistical program—it shows how using a statistical program can greatly enhance our understanding of the book's methods.

5 References

Allison, P. D. 1999. Comparing logit and probit coefficients across groups. *Sociological Methods and Research* 28: 186–208.

Fu, V. 1998. sg88: Estimating generalized ordered logit models. *Stata Technical Bulletin* 44: 27–30. Reprinted in *Stata Technical Bulletin Reprints*, vol. 8, pp. 160–164. College Station, TX: Stata Press.

Jann, B. 2005. Making regression tables from stored estimates. *Stata Journal* 5: 288–308.

Long, J. S. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.

Long, J. S., and J. Freese. 2003. *Regression Models for Categorical Dependent Variables Using Stata*. Rev. ed. College Station, TX: Stata Press.

———. 2006. *Regression Models for Categorical Dependent Variables Using Stata*. 2nd ed. College Station, TX: Stata Press.

Schumm, L. P. 2005. Review of Data Analysis Using Stata by Kohler and Kreuter. *Stata Journal* 5: 594–600.

Williams, R. 2006. Generalized ordered logit/partial proportional odds models for ordinal dependent variables. *Stata Journal* 6: 58–82.

About the author

Richard Williams is associate professor and a former chairman of the Department of Sociology at the University of Notre Dame. His teaching and research interests include methods and statistics, demography, and urban sociology. His work has appeared in the *Stata Journal*, *American Sociological Review*, *Social Problems*, *Demography*, *Sociology of Education*, *Journal of Urban Affairs*, *Cityscape*, *Journal of Marriage and the Family*, and *Sociological Methods and Research*. His current research, which has been funded by grants from the Department of Housing and Urban Development and the National Science Foundation, focuses on the causes and consequences of inequality in American home ownership. He is a frequent contributor to Statalist. His `gologit2` and `oglm` programs were inspired, in part, by his work with the earlier edition of Long and Freese's book.