The Stata Journal (2002)
2, Number 1, pp. 103–105

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

John Hendrickx
Wageningen University, The Netherlands
john_hendrickx@yahoo.com

Abstract. The new book Long and Freese (2001) is reviewed.

Keywords: gn0002, categorical data, regression models

1 Introduction

Long and Freese (2001) is similar in subject matter and structure to Scott Long's 1997 book (Long 1997). In the present book, however, Long and Freese concentrate mainly on the uses of Stata, in conjunction with their SPost package, for the estimation and interpretation of models with categorical dependent variables. A further difference is that, as the title suggests, limited dependent variables are beyond the scope of this book. Each chapter is preceded by a brief explanation of the statistical issues but lacks the depth of discussion in Long (1997). A large part of the book is devoted to post-estimation analysis using the SPost package. A basic knowledge of linear regression and statistics is assumed, but the brief statistical explanations of each model make it quite suitable for beginners in the area of categorical models. The SPost package and guidelines for interpreting categorical models make it an important book for all Stata users.

The book contains two parts, plus appendices. Part I comprises chapters 1 through 3 and deals with general issues relating to Stata and the SPost package. Chapter 2 is a particularly useful introduction to Stata and the main issues in analyzing data. My only comment about Chapter 2 is that user-written programs might have been highlighted a bit more. The relative ease with which Stata's capabilities can be extended and the availability of user-written programs through Stata's network capabilities are important advantages of Stata that new users should be aware of.

I found Chapter 3, which covers the general issues of estimation, testing, fit, and interpretation to be somewhat out of place. An explanation of the more general features of Stata output for ML models would have fit in better in Chapter 2, whereas the general discussion of the SPost commands could have preceded Appendix A on the syntax of SPost commands. The separate SPost programs and their uses are explained in subsequent chapters in the context of specific models and using examples, which will be much more useful to most readers. The general discussion in Chapter 3 is too abstract at this point.

Chapters 4 through 8 compose Part II, and discuss models for specific kinds of outcomes. Chapters 4 through 7 look in turn at models for binary outcomes, ordinal outcomes, nominal outcomes, and count outcomes. Each chapter begins with a short outline of the statistical aspects of the model. This will be adequate as a refresher or as an introduction for beginners, but researchers interested in the statistical intricacies should consult a second text such as Long (1997) for more in-depth information. The chapter continues with the estimation in Stata of the models in question and the interpretation of the output. This is followed by a discussion of tests on individual and multiple coefficients, measures of fit, and interpretation using predicted values. Each of these chapters also contains one or two paragraphs relating to the specific technique in question. Chapter 8 wraps up the book with a discussion of some general issues in regression models such as categorical independent variables, interactions, and nonlinear effects.

As already noted, post-estimation analysis of the results receives considerable attention in this book. Long and Freese reason that models for categorical dependent variables are nonlinear, and that interpretation must therefore focus on predicted values and marginal changes in the predicted values, given certain values on the remaining independent variables. The SPost package makes it a simple matter to make these calculations and present the results either as tables or as graphs. This can greatly facilitate the interpretation if the intended audience prefers it in terms of the untransformed dependent variables.

However, an interpretation in terms of the untransformed dependent variable is not quite as essential as Long and Freese suggest. Many people are comfortable instead with an interpretation of the linear predictors. While it is true that changes in the independent variable do not correspond to fixed changes in the dependent variable, they do correspond to fixed changes in a transformation of the dependent variable. A positive coefficient in (for example) a logit model means that higher values of that variable lead to an increase in the predicted probability, just not by a fixed amount. In many situations, a loose interpretation in this fashion will be adequate and less time-consuming than a full post-analysis in terms of predicted probabilities and marginal changes.

Although their book is intended for Stata users, I wonder if Long and Freese aren't overly loyal when they recommend Stata for graphing as well as statistical analyses. Excel and similar spreadsheet and graphing programs will often be quicker and more flexible, and will make it easier to export results to reports and presentations.

I was rather puzzled by Long and Freese's treatment of the numerous pseudo-R^2 measures. As in Long (1997), the formulas of these measures are duly noted, but no guidelines are given as to which might be preferable. I got the impression that Long and Freese have included these measures with some misgivings, to be used at one's own risk. Indeed, they note that there is no real relationship between these measures and optimal properties of the model. While this may be true, some guidelines could surely have been given here. A suggestion for future versions of the `fitstat` program is to allow users to select which measures are to be printed through the use of global macros. Most of

the time, researchers are interested in only one or two measures, so the remainder only confuse the picture.

2 Summary

These remarks are of course colored by my own opinions and preferences. On the whole, I found Long and Freese's book useful and informative. The book is clearly written and well structured. The treatment of the statistical issues is short but clear and will help bring beginners or those in need of a brief refresher up to speed. The discussion of post-estimation in conjunction with the SPost package will be helpful to beginners and experts alike, although I wouldn't go as far as to say that it will make post-estimation analysis enjoyable as the authors say on page 3. *Regression Models for Categorical Dependent Variables using Stata* is an essential book for Stata users interested in categorical data analysis.

3 References

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

Long, J. S. and J. Freese. 2001. *Regression Models for Categorical Dependent Variables using Stata*. College Station, TX: Stata Press.

About the Author

John Hendrickx is an assistant professor at the Management Studies Group of Wageningen University, the Netherlands. His research interests include social stratification, labor-market studies, and statistical methodology, particularly with respect to categorical data. He is the author of the widely used `desmat` package, among other Stata programs.