BOOK REVIEW

Regression Models for Categorical Dependent Variables Using Stata

JOSUÉ GUZMÁN, PhD

The new book by Long and Freese (1) describes how to estimate, test, fit, and interpret nonlinear regression models for categorical outcomes using the statistical package *Stata*™. It also includes *SPost*, which is an extension of post-estimation commands written by the authors. The book is complementary to another from the first author on the same subject (2).

Models for categorical dependent variables use nonlinear regression methodology where the outcome of interest is not continuous, and classical regression cannot be applied. They are special cases of generalized linear models (3, 4). The dependent variable can be binary (e.g., yes, no), ordinal (e.g., tumor grade), nominal (e.g., marital status), or a count variable (e.g., number of epileptic seizures). Explanatory variables can be either continuous or categorical, including interactions. Models for binary outcomes include: logit, probit, complementary log-log, and conditional regression (5). Ordinal regression includes ordered logit, ordered probit, continuation ratio, and generalized ordered logit models. Multinomial modeling is an extension of binary regression where interest is on the analysis of three or more unordered outcomes; a reference outcome is selected for comparative purposes. Models for count outcomes include: Poisson and negative binomial regression, and zero-inflated regression (6). Throughout these models, interest is not only on the partial effects of each covariate, but also on the ratio of odds.

Book Outline

The book is about estimation, testing, and prediction of models for categorical dependent outcomes using *Stata*, as well as on the interpretation of results. It is divided in two parts: I. *General Information*, and II. *Models for Specific Kinds of Outcomes*. Part I include the following chapters: 1. *Introduction*, 2. *Introduction to Stata*, 3. *Estimation, testing, fit, and interpretation*. Part II contains the following chapters: 4. *Models for binary outcomes*, 5. *Models for ordinal outcomes*, 6. *Models for nominal outcomes*, 7. *Models for count outcomes*, and 8. *Additional topics*. For each of the specific models, the authors describe the statistical model, and then proceed to maximum likelihood estimation using *Stata* (7), hypothesis testing (Wald tests, likelihood ratio tests), measures of fit, and interpretation. The authors describe the pertinent *Stata* commands, plus the *SPost* commands written by them. Additional topics include: ordinal and nominal covariates, interactions, nonlinear models, extending *SPost* to other estimation commands, and using *Stata* more efficiently. Finally, the book contains two appendixes: A. *Syntax for SPost commands*, and B. *Description of data sets*. Throughout the book, the reader is referred to some useful web pages including www.indiana.edu/~js1650/spost.htm, where other details about post-estimation commands and data sets used can be obtained.

Summary

The reviewer considers convenient the use of this book together with the often-cited Long book (2). After introducing the basic concepts on nonlinear regression and about *Stata*, the authors proceed to the main issues of maximum likelihood estimation, post-estimation analysis, testing (Wald and likelihood ratio tests) measures of fit using *fitstat* and *SPost* commands, and interpretation. For prediction purposes, in addition to *Stata*'s *predict* command, the authors include several commands on predicted value $\hat{y} = x\hat{\beta}$, predicted probabilities $\widehat{\Pr}(y = k)$, and predicted count or rate. Post-estimation commands *prvalue*, *prchange*, *prtab*, and *prgen* were written by the authors in order to compute specific predictions useful for the interpretation of categorical and count outcomes models. The new *Stata* 7 post-estimation command *mfx* to display model results in terms of marginal effects, which can be displayed as either derivatives or elasticities, is also discussed. The authors illustrate the use of different models with data sets on: labor force participation,

From the Department of de Biostatistics and Epidemiology, Graduate School of Public Health, Medical Sciences Campus, University of Puerto Rico, San Juan Puerto Rico.

Address correspondence to: Dr. Josué Guzmán, Department of de Biostatistics and Epidemiology, Graduate School of Public Health, Medical Sciences Campus, University of Puerto Rico, San Juan PR 00936-5067. Tel. (787) 758 2525 Ext. 1428, Fax: (787) 764 5831, E-mail: jguzman@rcm.upr.edu

scientific productivity, General Social Survey, careers of biochemists, and travel mode choice. Since a lot of data for categorical dependent variable regression comes from survey studies using sampling weights, clustering and stratification [and *Stata* contains survey (*svy*) estimation procedures for categorical dependent outcomes (*svylogit*, *svyprobit*, *svymlogit*, *svyologit*, *svyoprobit*, and *svycount*)] the reviewer considers ideal that *SPost* commands be extended and be applied to survey data generated using weights, clusters and strata. These would be similar to *stpm* commands developed for survival analysis using *Stata*'s *st* by Royston (8). Overall, however, the book is an excellent and useful reference for both, academic, biomedical and biosocial research purposes on categorical dependent (and count) variables via regression modeling and analysis. It offers good advises not only on the estimation of odds ratios and predicted probabilities, based on the pertinent model, but also emphasizes the important issue of interpretation of relevant results.

Resumen

El libro de Long y Freese (1) describe cómo estimar, hacer pruebas de hipótesis, fijar e interpretar modelos no-lineales para resultados categóricos usando el programa estadístico *Stata*™. Contiene además, los comandos denominados *SPost*, los cuales fueron creados por los autores para ser utilizados en post-estimación. El libro es complementario a otro del primer autor (2).

References

1. Long, JS, Freese J. Regression models for categorical dependent variables using Stata. College Station, TX: Stata Press; 2001.
2. Long, JS. Regression models for categorical and limited dependent variables. Thousand Oaks, CA: Sage Publications; 1997.
3. McCullagh P, Nelder J. Generalized linear models. 2d Ed. New York: Chapman and Hall; 1989.
4. Hardin J, Hilbe J. Generalized linear models and extensions. College Station, TX: Stata Press; 2001.
5. Hosmer DW, Lemeshow S. Applied logistic regression. 2d Ed. New York: John Wiley & Sons; 2000.
6. Cameron AC, Trivedi PK. Regression analysis of count data. New York: Cambridge University Press; 1998.
7. Gould W and Sribney W. Maximum likelihood estimation with Stata. College Station, TX: Stata Press; 1999.
8. Royston P. Flexible alternatives to the Cox model. Stata J 2001; 1: 1-28.