The Stata Journal (2008)
8, Number 1, pp. 122–128

Review of Applied Health Economics by Jones, Rice, Bago d'Uva, and Balia

Stephen P. Jenkins
Institute for Social and Economic Research
University of Essex
Colchester, UK
stephenj@essex.ac.uk

Abstract. This article reviews *Applied Health Economics* by Jones, Rice, Bago d'Uva, and Balia.

Keywords: gn0038, health economics, Stata texts

1 Introduction

Jones et al. (2007) provide an excellent introduction to the methods used by health economists for the statistical analysis of survey data. The book is not about health economics concepts or about econometric principles but, instead, about the combination of the two. It provides a practical guide to how to do applied research: it might be more accurately titled *Applied Health Econometrics*. Notwithstanding the health focus, the book will be a useful handbook for advanced undergraduates, graduate students, and researchers in many fields in addition to health.

The authors have well-established reputations in health economics research and teaching. They are all associated with the leading center for health economics in the United Kingdom (at the University of York) and have published in leading health economics and statistics journals using the methods described. Another benefit is that the material has already been "road tested": it is based on their short course and revised in the light of feedback received. The case studies are largely based on research published by the authors.

2 Coverage

The authors group their chapters according to the nature of their outcome variable and the types of survey data available. Thus chapter 1 introduces the five cross-sectional and panel datasets that are used in the case studies in the subsequent chapters: eight waves from the British Household Panel Survey (BHPS), cross-sectional and follow-up data from the British Health and Lifestyle Survey (HALS), cross-sectional data from wave 1 of the Canadian National Population Health Survey (NPHS), cross-sectional data from the World Health Organization Multi-Country Survey for the Indian state of Andhra Pradesh (WHO-MCS), and data from waves 2–4 of the Portuguese component of the European Community Household Panel (ECHP). Chapters 2 and 3 continue the

data description theme. The remaining three parts of the book consider categorical data (chapters 4 and 5), survival data (6 and 7), and panel data (8–11). Each chapter contains a case study analyzing one or more health outcome variables with empirical illustrations that are entirely Stata based. (Familiarity with basic Stata structure and syntax is assumed.) Indeed, the text contains substantial amounts of verbatim do-file and log file text (in Courier font to distinguish it from surrounding commentary). All the do-file code presented in the text is downloadable from the authors' web site, http://www.york.ac.uk/res/herc/hedg.html, together with the BHPS and HALS data files used in the analysis. The version of Stata used is not mentioned, but it appears to be Stata 8.

Table 1 shows a more detailed summary of the topics and methods covered, with chapter-by-chapter classifications of the dataset used, technique illustrated, and the corresponding Stata tools used. The table shows that the book's emphasis is on multivariable regression models of the many types most common among microeconometricians. And so, for instance, probit models and their generalizations are favored over logit-type models. The panel and survival data models allow for random intercepts but not random slopes (`xtreg`, `xtprobit`, `xtcloglog`, `pgmhaz8`, `reprob`). Multilevel models, popular among some noneconomics social science disciplines and which might be estimated by using `gllamm` or (in Stata 10) `xtmixed`, `xtmelogit`, or `xtmepoisson`, are not considered. On the other hand, applied econometrics books rarely discuss issues of nonresponse (panel attrition), how to test for attrition bias, or how to apply inverse-probability weighted estimators; this book does. Another feature I like is that the book does not simply fit models; it also shows how to undertake specification tests and how to draw out the implications of parameter estimates.

(*Continued on next page*)

Table 1. Topics and methods covered in *Applied Health Economics*

Chapter	Dataset	Topic	Stata commands and tools
Part I: Data description			
1. Data and survey design		Introduction to datasets and variables	
2. Describing the dynamics of health	BHPS	Data description	`do`, `log`, `iis`, `tis`, `generate`, `replace`, `recode`, `label`, `egen`, `sort`, `by`, `xtile`, `cumul`, `graph bar`, `graph export`, `tabulate`, `summarize`
3. Inequality in health utility and self-assessed health	NPHS	Description of distributions and basic regression methods	`graph twoway`, `centile`, `_pctile`, `table`, `glcurve`[a], `kdensity`, `global`, `sktest`, `regress`, `oprobit`, `intreg`, `predict`, `test`
Part II: Categorical data			
4. Bias in self-reported data	WHO-MCS	Generalized ordered probit models	`reshape`, `program`, `ml`, `gop`[b], `hopit`[b], `matrix`
5. Health and lifestyles	HALS	Regression models for multiple binary outcomes	`label`, `tabulate`, `describe`, `foreach`, `pcorr`, `icd9`, `fitstat`[a], `probit`, `mvprobit`[a], `mvppred`[a], `meffcon`[b], `meffdum`[b]
Part III: Survival data			
6. Smoking and mortality	HALS	Survival analysis with continuous time duration data	`quietly`, date functions, `list`, `count`, `stset`, `stsum`, `stdes`, `sts graph`, `graph combine`, `streg`, `stcurve`, `stsgen`, `estimates store`
7. Health and retirement	BHPS	Survival analysis with discrete time duration data	`forvalues`, `xtdes`, `ltable`, `pgmhaz8`[a], `cloglog`, `xtloglog`
Part IV: Panel data			
8. Health and wages	BHPS	Linear panel data models	`local`, `while`, `regress`, `xtreg`, `hausman`, `xthtaylor`
9. Modelling the dynamics of health	BHPS	Limited dependent variable panel models	`dprobit`, `xtprobit`, `quadchk`, `clogit`, `reprob`[a]
10. Non-response and attrition	BHPS	Testing for attrition inverse-probability bias; weighted estimators	Same as chapter 9
11. Models for health-care use	ECHP	Count data models	`nbreg`, `gnbreg`, `zip`, `zinb`, `ztnb`, `lcnb2_pan`[b], `lc_hurdle_pan`[b]

Note: Stata commands and tools are listed only once, in the chapter in which first used.
[a] User-written program available from SSC and/or *Stata Journal* archives.
[b] Program written by the authors with code shown in the text (also downloadable from their web site).

Some readers might be bemused by the classification of linear regression (`regress`), ordered probit regression (`oprobit`), and interval-censored regression (`intreg`) as part of the bundle of tools for data description rather than modeling. On the other hand, contrary to some social scientists' prejudices about econometricians' obsessions, there is also much space given to nonregression-based numerical and graphical methods for data description and data checking and cleaning. This forms the bulk of chapters 2 and 3. I also like the introduction to basic elements of Stata programming. Throughout the book, the authors show how tools such as `local` and `global` macros and `foreach` and `forvalues` loop constructions make analysis more effective. This could have been flagged more as a feature.

Reflecting the fact that the book is an offshoot from a real-life research program, the Stata tools used are a mixture of built-in commands and other tools and user-written commands. Taken all together, the portfolio used illustrates two important strengths of Stata, namely, its extensive suite of canned routines and its extensibility. Interestingly, none of the user-written programs drawn on here that are downloadable from SSC or the *Stata Journal* archives were originally written for analysis of health survey data, but clearly they are useful in this context. The authors themselves also develop several special maximum likelihood estimators with `program` and `ml` code, notably for generalized ordered probit models (allowing for nonparallel cutpoints, with heterogeneity, i.e., "hopit" models) and for panel-data count models (latent-class hurdle and negative binomial models). They also show how to derive average partial effects after multivariate probit estimation. So, not only can readers learn more about advanced Stata programming from the commentaries on the construction of the code, but they may also apply the econometric techniques in other contexts. The code is in the public domain.

3 Discussion

The book meets its stated aims well. But having had my appetite whetted for this sort of material for teaching and training, I would like more. A second edition might address several matters.

The book could benefit from greater overall editorial control providing greater cross-chapter consistency in style—both in the text and in the Stata programming. For example, some chapters use global macros to hold sets of covariate names, whereas others use local macros for the same task. Both methods work, but there is no explanation of the difference. I have a few other minor quibbles with the code used. For example, `tis` and `iis` are used instead of `tsset`, which was introduced in version 7. `stsplit` is cited as the means for episode splitting when preparing discrete-time survival data for analysis. It works, but in my experience, its use misleads students who become confused about the differences between estimation of continuous-time and discrete-time survival models. For the latter, I prefer to avoid all reference to st commands and implement episode splitting by using `expand`.

The book contains inconsistent references to user-written programs and how to obtain them. Chapter 2 uses `glcurve` to draw Lorenz and concentration curves, but

the chapter does not name the authors (Jenkins and Van Kerm 1999, 2007) nor, more importantly for readers, does it explain where to download the command (the latest version is a *Software Update* to *Stata Journal* volume 7, number 2). Chapter 3 does cite a *Stata Journal* article about `mvprobit` and `mvppred` (Cappellari and Jenkins 2003), but only several pages later is the `ssc` command referred to as a means to obtain them. And in the intervening pages, the `findit` command is cited as the means to obtain the `fitstat` command, but the authors are not mentioned. (They are Long and Freese [2006], who have now incorporated the `fitstat` command into their `spost` package.) The source for Mark Stewart's `reprob` command for dynamic binary dependent variable panel regression is cited as an unpublished paper. `findit` will not find it. In fact, the command has become the more powerful `redpace` (Stewart 2006), available from the *Stata Journal* archives.

Of course I am not complaining about the use of commands that I have co-written! The authors could have explained in one place (e.g., the *Preface*) the several ways to obtain user-written commands and how to get the latest version. Then they could have referred back to this whenever required later in the book. The *Preface* would also have been the place to cite the version of Stata being used and the implications of using later versions. A more systematic introduction to the use of do-files and log files would have been well-placed there, too.

The discussion of datasets (chapter 1) should be clearer about the relationship between the original datasets and the ones that are actually used later in the book. For example, it is helpful to know from the start that only the first 8 waves of the BHPS are used (15 are now available to registered users from the UK Data Archive). Similarly, the discussion of the ECHP is general, referring to all 14 participating countries and up to 8 waves of data (the maximum); in chapter 11, we finally learn that the case study is based on only 3 waves of data from Portugal. The authors could also be clearer about how users might obtain the Stata datasets used in their case studies. Five different surveys are discussed in chapter 1, and readers may gain the impression that Stata datasets based on them are also available. In fact, only data from the BHPS and HALS are downloadable from the authors' web site. Two is better than none, but readers would benefit from knowing precisely how to obtain the other datasets.

A related and more substantial point is that one important stage of analysis is rather downplayed in the book. There is a useful distinction between the dataset(s) that is available from a data distributor and an analysis dataset that is derived from it. For example, each wave of BHPS data consists of more than 10 files, each of which keys on different identifiers (distinguishing respondents, enumerated individuals, households, spells, income sources, etc.) and which contains different types of variables. To construct a panel dataset for analysis, one usually has to merge data from different files within each wave, and then append or merge files across different waves (depending on whether one wants long- or wide-format data). In my experience, it is this stage of analysis that is the most problematic for researchers (even BHPS experts) and most likely to lead to errors. The authors use one `bhps.dta` file for analysis, with no discussion of how it was created from the original files. Similar remarks apply to the other datasets. Even if the book itself has no space for the do-file code showing how to create the analysis datasets,

it would be useful to have this information on the authors' web site alongside the other code.

Although the authors emphasize their focus on hands-on empirical analysis, eschewing discussion of the microeconometric methods, I think the book would benefit from more systematic referencing of such discussions elsewhere. Most chapters have a final short *Overview* section, which is little more than a brief recap of the topics covered. This section would be more useful if it included "further reading"—citations to the relevant chapters in econometric and statistics texts and perhaps also to related health economics literature.

The book's *Introduction* could also be expanded to provide readers with a greater awareness of the microeconometric and health economics topics that the book does not cover. I am not a health economist and so am poorly placed to make suggestions about this. Nonetheless, I observe that all the case studies here refer to individuals and their health. None of the studies considers data about health providers (e.g., hospitals or doctors) or health insurers and which of these may have provided a greater role in the discipline of health economics outside Europe. On the methods side, I am aware that there is a substantial microeconometric literature within health economics that has debated the relative merits of the "Heckman selection" model and the "two-part" model for modeling data comprising a significant minority of observations with zero value for the outcome variable (e.g., health care expenditure) and the remainder with a positive value. (See Jones 2000 for a review.) Although the book considers only empirical research on topics that the authors themselves have undertaken, and it is an impressively wide range, it would be useful for readers to get a feel for what topics and methods have not been covered.

The text is remarkably free of typographical errors. One suggestion for the future is that the bullet points used to preface do-file command lines be omitted; they are redundant.

4 Conclusions

Applied Health Economics is a good practical guide to the application of microeconometric methods to survey data on health and related variables. It could be used as a complementary text for econometrics courses at the advanced undergraduate or postgraduate level or for specialist training courses, and it will be a useful reference for applied researchers in health and other fields.

5 References

Cappellari, L., and S. P. Jenkins. 2003. Multivariate probit regression using simulated maximum likelihood. *Stata Journal* 3: 278–294.

Jenkins, S. P., and P. Van Kerm. 1999. sg107: Generalized Lorenz curves and related graphs. *Stata Technical Bulletin* 48: 25–29. Reprinted in *Stata Technical Bulletin Reprints*, vol. 8, pp. 274–278. College Station, TX: Stata Press.

———. 2007. Software Update: sg107: Generalized Lorenz curves and related graphs. *Stata Journal* 7: 280.

Jones, A. M. 2000. Health Econometrics. In *Handbook of Health Economics*, ed. A. J. Culyer and J. P. Newhouse, 265–344. Amsterdam: North Holland.

Jones, A. M., N. Rice, T. Bago d'Uva, and S. Balia. 2007. *Applied Health Economics*. London: Routledge.

Long, J. S., and J. Freese. 2006. *Regression Models for Categorical Dependent Variables Using Stata*. 2nd ed. College Station, TX: Stata Press.

Stewart, M. B. 2006. Maximum simulated likelihood estimation of random effects dynamic probit models with autocorrelated errors. *Stata Journal* 6: 256–272.

About the author

Stephen Jenkins is the Director of the Institute for Social and Economic Research, University of Essex, UK (the home of the British Household Panel Survey), a research professor at DIW Berlin, and an associate editor of the *Stata Journal*. He is an applied economist who has contributed several often-downloaded commands to the SSC archive.