The Stata Journal (2005)
5, Number 2, pp. 274–278

Review of Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models by Vittinghoff, Glidden, Shiboski, and McCulloch

Stanley Lemeshow
The Ohio State University
lemeshow.1@osu.edu

Melvin L. Moeschberger
The Ohio State University
moeschberger.1@osu.edu

Abstract. The new book by Vittinghoff et al. (2005) is reviewed.

Keywords: gn0028, linear regression, logistic regression, survival analysis, repeated measures, generalized linear models, complex surveys

1 Introduction

Finding the right book for an intermediate level biostatistics course is not easy. Here at the Ohio State University School of Public Health, we teach a year-long sequence of three courses in applied statistical methods that graduate students take during their first year. The last of these quarter-long courses is composed of a set of topics that we believe students need to know something about before graduating. These topics include repeated-measures analysis of variance, linear regression, logistic regression, and survival analysis. Many of these topics have corresponding full quarter-long courses that students can take the following year, but for some students, especially those not majoring in biostatistics, this may be the last course in biostatistics they ever take. For other students, a topic in this course may tweak their interest to take the full course later on. Up to this point, we have never had a required text for our course since we were not aware of one that sufficiently developed each of the topics we cover. After reviewing this new book by Vittinghoff, Glidden, Shiboski, and McCulloch, we feel that we have found a text that we could try in our course and that students would find to be a valuable supplement to the material we teach.

The approach we take in our course is similar to the one used in this book. We emphasize the application of the statistical methods and focus on the use of statistical software to apply the methods to real data. Our students are drilled on proper interpretation of statistical software output. The authors have chosen to present computer output from Stata to illustrate each of the methods. This is the package we most commonly use in our teaching, as well, so again, this is a good match.

As prerequisites for the book, the authors suggest that students be familiar with "the basic tools: paired and independent samples t tests, simple linear regression and one-way ANOVA, contingency tables and χ^2 (chi-square) analysis, Kaplan–Meier curves, and

the logrank test" (page vii). All of these topics are covered in the two quarters before the course we would like to use this book for, except Kaplan–Meier curves and the logrank test. We teach these in our introduction to survival analysis in this course.

2 Summary

The book is composed of eleven chapters. Chapter 1 introduces the family of multipredictor regression methods. Using a dataset, the authors motivate the reader to understand that the methods they have learned up to this point are not enough to deal with measurements taken over time on the same subject, the clustering of subjects within physician practices, and the need to control for confounding.

Chapters 2 and 3 provide a review of basic topics. These include descriptive statistics, t tests, one-way analysis of variance, linear regression, contingency tables, survival analysis (Kaplan–Meier curves and logrank tests), and bootstrap confidence intervals. Students would not want to be learning these methods for the first time from the roughly 60 pages that cover all of these topics. However, the high points are hit, and students with a solid foundation should find these pages to be a good review. For the most part, the authors present a good treatment of descriptive and graphical statistics. However, we believe it is very dangerous to suggest that "we may treat a categorical variable as a numerical score", as the authors suggest (page 8). Noticeably absent is a discussion of stem and leaf plots, an important topic in exploratory analyses. Otherwise, the chapter is good and very useful.

Chapter 4 focuses on regression beyond a single predictor. This is a solid chapter. We feel that the list of additional references is far too thin. There are many excellent, easy-to-read books that could have been referenced in this chapter. There are a few topics that are not fully developed. For example, the discussion of multiple comparisons is very limited. There is a noticeable absence of testing using the generalized F-statistic.

Chapter 5 discusses strategies for selecting predictors to be included in models. We find the topics covered in this chapter to be important, but the discussion of the topics is very thin, perhaps to the point of limiting the usefulness of this chapter. For instance, classification and regression trees are important research tools for many researchers, but this topic receives only 20 lines of discussion in the book.

Chapter 6 devotes 40 pages to a discussion of logistic regression. In general, the topics covered are important and appropriate. Again, 40 pages is not nearly enough to cover many of the topics in enough detail to provide students with the tools they need to independently apply a particular technique. For example, the section on checking model assumptions might be particularly frustrating for students. There is no indication in that section of the book of how to get Stata to produce the diagnostic statistics, and the discussion of how to interpret the plots is too sketchy to be very useful. However, these are issues that can be covered by course instructors, and we recognize that the trade-off for covering many topics is to cover many of them in minimal depth.

Chapter 7 devotes about 40 pages to survival analysis, particularly the Cox proportional hazards model. The material in this chapter, together with the material covered in Chapter 3, is a rough approximation of what we think students should know if this were the only course they would ever have on survival analysis. We would have preferred that the references would have mentioned some of the other books that are available and readable to allow students to take the next steps in this area beyond what is covered in this book. In general, however, this chapter presents a nice treatment of survival analysis. Two topics get a rather cursory treatment (one page each), namely, truncation and interval censoring. Many studies involve both of these concepts. For example, in the Framingham Heart Study, subjects entered the study at different ages and thus have delayed entry (left truncation). This conditional entry time must be used in the analysis to get unbiased estimates of the survival function. Also, in this study, the events of interest (excluding death, of course) are frequently only known to have occurred between successive exams. That is, the events are known only to have occurred in an interval (the classic case of interval censoring). The authors treat graphical techniques and the Cox proportional hazard and its refinements adequately.

Chapter 8 deals briefly with a number of important topics. These include repeated-measures analysis of variance, hierarchical data, longitudinal data, generalized estimating equations (GEE), and random-effects models. Davis (2002) is an important reference. Bootstrap methods for constructing confidence intervals are also covered. Obviously, this is an awful lot of material to cover in 35 pages, and students will need significant support to master any of these methods. They will not be able to do it from this book alone.

Chapter 9 presents a discussion of generalized linear models. The chapter is quite short and features a few examples of Stata's `glm` output. Chapter 10 provides fewer than 10 pages dealing with the analysis of complex sample surveys. Stata is, of course, a very valuable program for analyzing survey data. Levy and Lemeshow (1999) present an extensive discussion of the use of Stata for complex surveys, yet this is not suggested as a reference for this chapter. It would be impossible to imagine that this chapter, as short as it is, could provide students with enough useful material to have them analyze survey data effectively.

3 Conclusion

This book covers the five generic problems that we cover in the third quarter of a year-long course, namely, ANOVA, repeated-measures ANOVA, ordinary least-squares linear regression, logistic regression, and survival analysis. ANOVA is covered very lightly, but that is not a problem since the students will have some exposure to this topic before entering the class. Some topics, such as multiple comparisons, can be added easily through supplemental readings. Repeated measures ANOVA coverage is adequate for an introductory treatment; however, we would need to supplement the text somewhat. Ordinary least-squares linear regression is covered in a fairly standard way and will be fine for the class. Again some topics, such as generalized F tests, can be added easily

through class notes. The logistic regression chapter is, perhaps, the strongest in the book, but we would need to supplement the material presented since many topics are covered too superficially. The survival analysis section is weak on interval censoring (a very important topic on which an entire book, as yet unpublished, is being written) and truncation. The discussion of cumulative incidence curves is somewhat useless. This is not a large problem since we do not cover these topics in this particular course. Since each of the five topics have entire books devoted to them, we would use those books as references for further reading and study. We would have been much more comfortable with this book had the references at the end of each chapter been a bit more inclusive.

In summary, we are very favorably impressed with this book as a text for our course. It gives an adequate introduction to the basic topics and will do fine as a central text. It covers all the important topics and attempts to integrate the regression concepts common to each of the methods. Even though we would supplement some topics with additional readings or class notes (a comment which applies to all texts), we will try this book out as a text for our course this year.

4 References

We have found the following references to be useful to students taking the course we teach at Ohio State. Perhaps in the next edition, more of these could be referenced by Vittinghoff, Glidden, Shiboski, and McCulloch.

Davis, C. S. 2002. *Statistical Methods for the Analysis of Repeated Measurements*. New York: Springer.

Dean, A. and D. Voss. 2000. *Design and Analysis of Experiments*. New York: Springer.

Hosmer, D. W., Jr. and S. Lemeshow. 1999. *Applied Survival Analysis: Regression Modeling of Time to Event Data*. New York: Wiley.

—. 2000. *Applied Logistic Regression*. 2nd ed. New York: Wiley.

Klein, J. P. and M. L. Moeschberger. 2003. *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd ed. New York: Springer.

Kleinbaum, D. G. 1994. *Logistic Regression: A Self-Learning Text*. New York: Springer.

—. 1996. *Survival Analysis: A Self-Learning Text*. New York: Springer.

Kleinbaum, D. G., L. L. Kupper, K. E. Muller, and A. Nizam. 1998. *Applied Regression Analysis and Multivariable Methods*. 3rd ed. Pacific Grove, CA: Duxbury.

Levy, P. S. and S. Lemeshow. 1999. *Sampling of Populations: Methods and Applications*. 3rd ed. New York: Wiley.

Lohr, S. L. 1999. *Sampling: Design and Analysis*. Pacific Grove, CA: Duxbury.

Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. 1996. *Applied Linear Statistical Models*. Boston: McGraw–Hill.

Selvin, S. 1995. *Practical Biostatistical Methods*. Pacific Grove, CA: Duxbury.

Vittinghoff, E., S. C. Shiboski, D. V. Glidden, and C. E. McCulloch. 2005. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. New York: Springer.

About the Authors

Stan Lemeshow is Dean of the School of Public Health and Director of the Center for Biostatistics and Mel Moeschberger is Professor and Interim Chair of the Division of Epidemiology and Biostatistics, both at The Ohio State University. Both have published textbooks in topics covered by the book under review. These topics include logistic regression, sampling, and survival analysis.