The Stata Journal (2003)
3, Number 2, pp. 203–207

Review of Statistical Modeling for Biomedical Researchers by Dupont

Joanne M. Garrett
University of North Carolina
joanne_garrett@med.unc.edu

Abstract. The new book by Dupont (2002) is reviewed.

Keywords: gn0007, biostatistics

1 General comments

Let me start with a disclaimer: the term "biomedical researcher" may include many different types of people doing research in the medical field. My own experience has been exclusively with physicians doing clinical research, so this review is based on how I think William Dupont's new book would be viewed by such an audience.

The preface states that it "will enable biomedical researchers to use a number of advanced statistical methods that have proven valuable in medical research". It further states that "[t]he emphasis is on understanding the assumptions underlying each method, using exploratory techniques to determine the most appropriate method, and presenting results in a way that will be readily understood by clinical colleagues" and finally that the goal "is to allow investigators to effectively use some of the most valuable multivariate methods without requiring an understanding of more than high school algebra. Much mathematical detail is avoided by focusing on the use of a specific statistical software package". Some of the topics covered include linear regression, logistic regression, survival analysis, and Poisson regression.

The author uses his book as the text for a second-level biostatistics course for a Master's of Public Health (MPH) degree program. Therefore, he suggests that readers should have taken a basic biostatistics course beforehand. He also says the book is "self-contained" but recommends introductory biostatistics and epidemiology texts to supplement the material. On reading this preface, I got very excited. At last, a book had been written that addresses the needs of the highly skilled but sometimes statistically challenged medical researchers who want to have a better understanding of the appropriate statistical methods to analyze their data, yet who have little training in traditional biostatistics. For the most part, such researchers are not particularly interested in the underlying statistical theory or formulas. What matters is how to apply the appropriate test to their data, to know the assumptions that make that test appropriate, and to understand when violating those assumptions may make a result invalid. Further, they need to be able to present results to their peers in an easily understandable way, while still maintaining the integrity of any conclusions. This is not to say that we expect medical researchers to eliminate the need for biostatisticians.

Ideally, a study is carried out by a research team of individuals, each contributing a unique expertise, among which statistical support is key. But with easy access to computers and simple (though powerful) programs like Stata, many researchers would like to explore their data themselves. The excitement of seeing that first statistic or graph on the screen after months or even years of data collection is very appealing. Additionally, there is the efficiency of the instantaneous ability to do the next step based on a previous result without a 2- or 3-week delay waiting until the statistician has time to respond to the request. In the best of all possible worlds, the statistician will be a co-investigator on the study, will have a good working knowledge of the science, and will have a vested interest (publications and academic advancement, i.e., a pay raise) in seeing it through to a superb conclusion. The reality is, however, that many medical investigators, particularly those early in their careers, do not have the luxury of ready access to good statistical support, while most statisticians are over-committed on many different projects, so that they do not have the luxury of becoming as well versed in the science of each study as they might like. It is a worthy goal to ease the burden of overworked statisticians and help assure that medical researchers are knowledgeable enough to carry out most analyses correctly. A book that promises to help fill this gap is definitely needed. In some ways the book lives up to its promise, but in other ways, it is more problematic.

The discussion of each type of statistical analysis (e.g., linear regression, survival analysis) starts with simple methods and then proceeds to more complicated versions of the same type of analysis. Some of the chapters might be easier to follow if the first section of each gave more review of when a type of analysis is most likely to be used, perhaps with some simple examples, before launching into the calculations of the tests and more complicated examples. The author does this briefly, but I think it could be expanded. This lets people know where they are going first and then tells them how to get there. Although the math is not complicated, the many formulas and hand calculations can cause the mind to wander and make readers wonder how much they must assimilate before they can appropriately carry out an analysis. One of the very appealing things about the organization of the Stata manuals is that they present examples from actual studies first, tell you how to get there, interpret the results, and finally present the underlying statistics that are being used.

The author does not intend the book to replace Stata documentation, and he assumes that the reader will have read the *Getting Started* manual. Each type of analysis is explained first with formulas. Then follow the Stata statements needed to apply the method and thorough explanations of those statements. The author also makes the datasets for all examples available from a web site. Some of the Stata code may not be as efficient as one might like. For instance, when demonstrating how to do some analyses on a subset of the data, `keep` is first employed to retain only the data for that subset. An alternative that would not change the data is simply to identify the subset with an `if` condition. There is abundant use of graphics, naturally an important way to convey study results, particularly to nonquantitative audiences. Unfortunately, the book was published too early to be able to exploit the new graphics commands in Stata 8. Readers will need to know how to alter the statements to reproduce the

graphs in the book if they are using the latest version of Stata, or the book will need to be updated to use the new commands. This is an ongoing problem for any book that uses computer code, although it seems unlikely that Stata will change its graphics again quite so extensively in the foreseeable future.

2 Chapter-by-chapter comments

Chapter 1: Introduction

The introductory chapter of the book covers some basic descriptive statistics, definitions, and a brief discussion of Stata. This chapter is a nice review, although I'm not sure why t-tests are included here and not other bivariate tests. For instance, if bivariate tests are to be included, it might make sense to briefly describe simple 2×2 tables and Pearson chi-square tests because many medical outcomes are binary, and this sets up the descriptive work that usually precedes more complicated modeling. Alternatively, the t-test section could be incorporated into the simple linear regression or analysis of variance chapter.

Chapters 2 & 3: Linear Regression (Simple and Multiple)

The next two chapters cover linear regression, describing basic theory of the simple model, calculating predicted values, testing some assumptions about the model, and transforming variables, if necessary. This is followed by a discussion of multiple linear regression. Many of the examples used throughout these chapters come from the Framingham Heart Study. The tests of assumptions suggest using natural log transformations on some of the variables, and the transformed variables are used from there on out. It would be useful to have some discussion on when the violation is severe enough to bias the results if you don't transform, and then on how to interpret results in transformed units. Medical researchers can relate to the values of systolic blood pressure, but the natural log of systolic blood pressure has no intuitive interpretation. It might have been better to give an example that could use variables in their original units for the main examples, and then give other examples later using transformed variables. The section on "Automatic methods in model selection" starts with the comment "When a large number of covariates are available, it can be useful to use an automatic selection program for this task." This statement may be misleading, particularly if the sample size is small. Unfortunately, it is common among medical researchers to dump every variable into the model and have the program come up with a result, valid or not. The goal of this book should be to help guide the reader in a strategy of model reduction that will yield a valid model. So, perhaps a bit more on model reduction strategies would be helpful.

Chapters 4 & 5: Logistic Regression (Simple and Multiple)

The next two chapters cover logistic regression, first giving background, and then using a model with one independent variable, and expanding to a model with several variables. These chapters were thorough with many useful explanations and examples. I was a bit puzzled why there was a section on 2×2 case-control studies in the middle

of the discussion, without analogous discussion about cohort or cross-section studies. The multiple logistic regression chapter starts with background using stratified analysis and then demonstrates how to do such analysis using logistic regression. The stratified analysis uses nominal variables with several categories, each requiring the use of dummy variables in the model. In fact, most of the examples use these dummy variable models, including interactions between them. It might have been better to begin with a simpler multiple-variable model, composed of interval and dichotomous variables, before launching into these messier models. Models with several multiple-level nominal variables are usually the exception. Also at the end of Chapter 5 is a discussion about modeling with missing values. Because this is relevant to any statistical model, perhaps a discussion on handling missing data should have come earlier in the text.

Chapters 6 & 7: Survival Analysis (Introduction and Hazard Regression Analysis)

The chapters on survival analysis were among the best in the book. The discussion followed a logical sequence from definitions and basic tests through more complicated modeling, all illustrated with examples. The author chose to use the term "relative risk" to describe the ratio of the two hazards (or instantaneous incidence rates) rather than the usual "hazard ratio". Also, although proportional hazards are discussed in adequate detail, there is very little on testing whether the data violate the assumption, what happens if this violation is ignored, or ways to correct for it in the analysis. There is a brief section on using stratified Cox models, but had I not known that this is one approach to get around violation of proportional hazards, I think I would not have figured it out as written. Saying stratification is a way to "weaken" the assumption may not be clear enough.

Chapters 8 & 9: Poisson Regression (Introduction and Multiple)

About the only places to find information on Poisson regression are in theoretically written text books, which provide much more than most people want to know, or in brief sections of books that cover many other topics. So an advantage to these two chapters is that Poisson regression is covered in enough detail to be thorough, while using examples along the way. It is a nice combination of theory and practice.

Chapters 10 & 11: Analysis of Variance (Fixed Effects and Repeated Measures)

The final two chapters cover analysis of variance. One might expect to find these chapters earlier in the book, but they can stand alone, so perhaps the placement isn't that important. The repeated measures chapter covers some standard techniques as well as an introduction to GEE modeling, all demonstrated by example.

3 Summary

A positive aspect to this book is that each set of formulas and many examples are followed with Stata statements and thorough explanations of them. This allows readers

to reproduce the results, and potentially to apply this knowledge to similar studies. In addition, there is a lot of emphasis on using graphics to illustrate the results. On the other hand, my definition of "self-contained" implies a book that a reader could use to learn what is needed even without other resources. I envision this text being useful as a supplement to classroom lectures, as the author is doing. However, I think it would be difficult for most medical researchers to proceed on their own without verbal clarification on some of the explanations as presented. In addition to being used as a classroom text, this book could be helpful as a reference for Master's level biostatisticians or epidemiologists who serve as research assistants to medical researchers.

4 References

Dupont, W. D. 2002. *Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data*. Cambridge: Cambridge University Press.

About the Author

Joanne Garrett is a Professor in the Department of Medicine, Associate Director of the Robert Wood Johnson Clinical Scholars program, and Adjunct Professor in the Department of Epidemiology in the School of Public Health at the University of North Carolina in Chapel Hill, North Carolina. Most of her research has been in health services research, including medical decision making and outcomes of care. For the last 17 years, she has taught biostatistics, quantitative epidemiology, and other research methods to physicians planning careers in clinical research.