The Stata Journal (2006)
6, Number 3, pp. 420–424

Review of A Gentle Introduction to Stata by Acock

Michael Mulcahy
Department of Sociology
University of Connecticut
Stamford, CT
michael.mulcahy@uconn.edu

Abstract. This article reviews *A Gentle Introduction to Stata* by Acock.

Keywords: gn0033, introductory statistics, social science, teaching Stata

1 Introduction

Alan C. Acock's *A Gentle Introduction to Stata* aims to introduce Stata to readers who have not only no prior knowledge of Stata but also no experience with any other statistical software and are just learning how to use statistics in the social sciences. This aim is reflected in Acock's extensive use of Stata's user-friendly menu system and dialog boxes as he moves the "true beginner to a level of competence using Stata" (xix) over 12 chapters. In contrast to much of the standard Stata documentation, which is organized around specific commands, this book follows a thematic progression related to preparing and executing basic social science statistical analyses.

The book features chapters devoted to creating and preparing new and existing datasets for analysis; generating univariate and bivariate descriptive statistics and graphs; and more advanced analyses such as bivariate categorical data analysis, various kinds of means tests, bivariate correlation and regression analysis, analysis of variance, and multiple and logistic regression. Each chapter combines general but highly accessible discussions of the subject matter with hands-on illustrations of the central procedures and analyses. The text prompts the reader through these applications and exercises using real data drawn from important existing social science datasets from the United States, such as the National Longitudinal Survey of Youth, the General Social Survey, or the High School and Beyond datasets, as well as a few datasets constructed by Acock to illustrate certain points efficiently. In addition to treating the central chapter themes, each chapter also contains information and instruction about statistical methods and Stata capabilities that go beyond the examples at hand. This approach gives the reader a thematically and substantively focused introduction to each chapter's subject matter, as well as a sense of some of the broader statistical and substantive issues and wider applications of Stata commands.

Each chapter concludes with a summary of the chapter's main points and practice exercises. Acock achieves impressive depth and scope in this introductory text, particularly given his commitment to assume no previous knowledge of statistical software on the part of the reader.

2 Contents

The first chapter introduces the reader to Stata screens and screen preference options, Stata dialog boxes and datasets, and the book's linguistic conventions. This chapter also moves directly to hands-on experience with Stata datasets, prompting to begin basic operations with an existing dataset on page 7. Acock describes accessing Stata datasets stored on the Internet. The second chapter focuses on procedures and techniques associated with creating a new dataset. The book shows the reader how to create a Stata dataset from the raw data gathered with a survey instrument. The chapter covers general aspects and Stata procedures for data coding; generating variable and value labels; and hand-entering, saving, and checking data.

Chapter 3 takes the reader through steps that are often necessary in preparing to use an existing dataset. Using data extracted from the U.S. Department of Labor's 1997 National Longitudinal Survey of Youth, this chapter takes the reader from accessing the dataset to creating a new scale variable based on four items from the original dataset. Along the way, the chapter addresses Stata dictionary files and codebooks; aspects of developing a research plan; describing, recoding, reverse-coding, modifying, and creating variables; and saving subsets of variables in separate datasets. This chapter also includes instructive sections on assessing and dealing with missing values in existing datasets. Chapter 4 introduces the reader more systematically to the basic structure of Stata commands, generating and saving do-files based on commands initially created through the dialog boxes of the menu system, and saving commands and results in log files. This chapter walks the reader through some basic commands for summarizing data and listing observations, adding value labels to lists of categorical variables, and generating pie charts of these variables. A major focus of the chapter, however, is on gathering commands like these into do-files, clarifying these commands by adding comments to those files, executing individual commands or sequences of separate commands from a do-file, and saving the do-files for future use as templates or for replicating results. Chapter 4 also demonstrates the alternative log-file approach to generating a permanent record of command sequences. The chapter also includes practical and helpful tips about transferring results from Stata to word processing programs.

Chapters 5–11 cover statistical analyses. Chapter 5 turns to univariate statistical descriptions and graphs. The first few pages of the chapter briefly introduce some basic concepts in descriptive statistics, such as levels of measurement, measures of central tendency (mode, median, mean), and measures of dispersion and shape (standard deviation, skewness, kurtosis). Later sections of the chapter introduce and illustrate Stata procedures for obtaining frequency distributions, horizontal and vertical bar charts, pie charts for unordered categorical variables, and summary statistical tables and histograms for ordered categorical variables. Box plots are added to histograms and summary descriptive tables in the chapter's discussion of interval-level variables. Along with the basic "how to" guidelines and examples, Acock discusses the different uses, strengths, and weaknesses of the descriptive statistics and summary techniques introduced. The examples and exercises there use various General Social Survey measures of sex, marital status, political views, education, and hours spent using the Internet. As in previous chapters, Acock discusses techniques for generating descriptive statistics and graphs by using both direct commands and Stata's menu-based dialog boxes.

Chapter 6 explores analyzing relationships between two categorical variables. Concepts and techniques covered include definitions of cross tabulations, dependent and independent variables, the Pearson chi-squared statistic, and degrees of freedom, among other topics. Using another dataset drawn from the General Social Survey, Acock demonstrates generating unordered and ordered cross tabulations of two categorical variables and of a quantitative and categorical variable, as well as bar charts relating a categorical and a quantitative variable. The chapter relies primarily on the dialog boxes of the Stata Statistics menu to generate these tables and charts. It also includes brief but helpful introductory discussions of other measures of association, such as Cramér's V, Kendall's tau-b, Goodman and Kruskal's gamma, and odds ratios. Acock also shows how to enter data into a table and thus calculate measures of association when you encounter a table with no percentages or measures of association. As a bonus, this chapter also demonstrates using Stata's `findit` command to access and download tables for chi-squared tests, F tests, and t tests and to display these tables in the Results window with simple one-word commands.

Chapter 7 discusses and introduces techniques for testing means. Acock nicely introduces the distinction between random sampling and randomization, and he shows Stata procedures for generating a wide assortment of means tests and comparisons. The chapter's means tests include z tests of proportions in one sample and two samples, t tests of means in one sample, two-sample t tests of group means (with equal and unequal variances), and repeated-measures t tests. A concluding section discusses nonparametric alternatives to the t tests, such as the Mann–Whitney two-sample rank-sum test and median tests. This chapter also invites the reader to gain more practice with do-files by posing the task of extensively recoding an income measure. Acock briefly discusses R^2 and Cohen's d as measures of the strength of differences between means, along with guidelines for computing these statistics by hand (Stata does not directly compute them) and interpreting the results. I found the chapter's coverage of power analysis in Stata to be particularly useful. This chapter also neatly highlights the real-life implications of the statistical and data-analytic issues.

Chapter 8 covers topics related to bivariate correlation and regression. The chapter starts with a discussion of scatterplots, illustrated with hands-on examples that use the dialog boxes, including some more advanced options. Acock then progresses through discussions and practice examples of regression line plots, casewise and pairwise deletion approaches to correlation analysis, regression analyses, rank-order correlations for ordinal data (Spearman's rho), standardized and unstandardized measures of reliability and internal consistency (Cronbach's alpha), and a measure of interrater agreement (Cohen's kappa).

In chapter 9, Acock introduces Stata procedures for the analysis of variance (ANOVA), including one- and two-way ANOVA, and the analysis of covariance (ANCOVA). The discussion of two-way ANOVA concludes by showing the reader how to generate a graph to display the results of the analysis. This chapter also resumes and extends the discussion of repeated-measures t tests begun in chapter 7 to cover repeated-measures ANOVA. Acock uses this opportunity to illustrate Stata's `reshape` command to transform Stata datasets from `wide` to `long` format. Chapter 9 briefly addresses the ρ_i measure of intraclass correlation.

Multiple regression analysis is the topic of chapter 10. Building on the discussions of bivariate regressions and ANOVA in the previous two chapters, it starts by demonstrating the dialog box and direct-command approaches to regression analysis in Stata. Acock patiently discusses each section of the regression analysis results, comparing the ANOVA in the regression table and the ANOVA analyses of chapter 9, as well as the relationships between the multiple regression R^2 and the r^2 statistic of the bivariate analyses in chapter 8. This chapter, the book's longest, also discusses standardized regression coefficients and using beta weights and increment in R^2 to compare and assess the magnitude of the individual variable effects, including directions for accessing user-written commands for analyses of the semipartial correlations and R^2. Chapter 10 also introduces a wide range of Stata's diagnostic tools and strategies for regression analysis: graphic and tabular procedures for examining the distribution of the dependent variable, approaches for identifying and dealing with outliers (e.g., Stata's `dfbeta` command, robust regression), techniques for creating scatterplots to identify heteroskedasticity, and commands for computing a variance inflation factor to check for multicollinearity among independent variables. This chapter also includes discussions and examples of categorical independent variables, nested regression modeling strategies, and a particularly useful introduction to creating interaction variables and interpreting interactions in regression analysis.

Chapter 11 covers logistic regression analysis, starting with a concise but lucid discussion, with graphic illustrations, of the logistic distribution and the problems arising from applying ordinary least-squares estimators. The discussion proceeds to odds ratios and the rationale for the logarithmic transformation of the odds ratio. The rest of the chapter illustrates logistic regression in Stata with data from the National Longitudinal Survey of Youth, 1997. The chapter covers both `logit` and `logistic`, with a careful discussion of interpreting the odds ratios in the output for `logistic`. This chapter also includes sections on testing hypotheses about individual parameters and sets of parameters, in addition to revisiting the previous discussion of nested modeling strategies with an explanation of extending Stata's `nestreg` command to logistic regression.

The last chapter provides a brief survey of available Internet resources, books, courses, and data sources for readers who want to continue developing their Stata skills.

3 Assessment

Although I do not fit the profile of the absolute Stata and statistics beginner for whom this book is primarily intended, I still found Acock's book instructive. Other readers in the early stages of their acquaintance with Stata will share that experience. At its best, the text moves effortlessly from lucid statements of the basic statistical issues to concrete Stata applications and exercises that let the reader learn by doing. That said, the challenges of attempting to start from true scratch are formidable and not always met with equal success. To be fair, I realize that Acock's stated aim is to introduce Stata to the reader who has no background in statistical software packages but is concurrently acquiring a first introduction to statistics elsewhere. Because of these divergent expectations, the text is sometimes uneven in the implicit assumptions made about prior statistical knowledge, and the progression through the statistical material is sometimes less rigorously systematic than is the progression through Stata commands and procedures. For example, chapter 5 discusses measures of skewness and kurtosis, but the systematic introduction of these statistics does not occur until chapter 10. This structure contrasts with the explicit definitions of the more basic concepts of nominal, ordinal, and interval variables in the same chapter. This is just one of several cases in which concepts are used unceremoniously before they are systematically introduced.

In other places, Acock's writing sometimes falls short of the clarity and avoidance of ambiguity that one would hope for in a truly introductory text. For example, in chapter 3 creating a scale variable is illustrated by combining four separate measures, each of which is a four-level categorical measure. Although such a potential overlap in referents may go unnoticed by readers with more background in Stata and statistical analysis, it could be avoided easily enough and might generate ambiguity for the genuine novice. Indeed, in other places Acock shows a sensitivity to just this kind of linguistic ambiguity. For example, in chapter 8 in discussing bivariate regression, he warns the reader not to confuse the beta weights currently under discussion with the beta coefficients of the same magnitude discussed in the preceding example.

On the whole, however, these are only minor shortcomings in Acock's overall successful effort to introduce Stata to readers with little background in statistics and no background in statistical software. The envisioned readers will find that, above all, Acock has kept his word by providing a dense but highly accessible introduction that will help the beginning Stata user to advance from zero to a command of a remarkable range of basic univariate and multivariate statistical analyses in Stata, while avoiding many of the frustrations and misinterpretations that dogged the path of those who learned Stata without the benefit of such a "gentle", yet authoritative, introduction.

4 References

Acock, A. C. 2006. *A Gentle Introduction to Stata*. College Station, TX: Stata Press.

About the author

Michael Mulcahy is a sociologist at the University of Connecticut who does comparative-historical and quantitative research in political economy, labor movements and other social movements, political sociology, organizations, and stratification.