The Stata Journal (2023)
23, Number 4, pp. 1074–1081 DOI: 10.1177/1536867X231212448

Review of Alan Acock's A Gentle Introduction to Stata, Revised Sixth Edition

Andrew Musau
Molde University College
Molde, Norway
amus@himolde.no

Abstract. In this article, I review *A Gentle Introduction to Stata*, *Revised Sixth Edition*, by Alan Acock (2023, Stata Press).

Keywords: gn0098, book review, introduction to Stata, teaching statistics, behavioral and social sciences, Stata

1 Introduction

I first encountered Alan Acock's *A Gentle Introduction to Stata* more than a decade ago while serving as a teaching assistant in a graduate-level course in econometrics. The instructor for the course had chosen the third edition of the book as one of two required textbooks, and the students' feedback on the book at the end of the semester was overwhelmingly positive. To date, the book has received more than 1,250 citations according to Google Scholar and is arguably the most successful introductory Stata book on the market. Previous reviews of the book in the *Stata Journal* include Mulcahy (2006) (first edition) and Collier (2015) (fourth edition).

It is not difficult to identify reasons for the immense success of Acock's book. He begins with the assumption that the reader has no prior experience with Stata or any other statistical software package. Therefore, there is extensive reliance on Stata's user-friendly menus and dialog boxes throughout the book instead of written commands and lines of code, which can appear daunting for new users. Acock takes a thorough approach to instructing the reader in all aspects of using Stata, including data management, cultivating good work habits (such as using do-files), exploring data with basic statistics (including graphical displays), and conducting analyses using standard statistical tools such as correlation, linear and logistic regression, and parametric and nonparametric tests of location and dispersion. Additionally, he presents more advanced topics like multiple imputation accessibly.

Acock emphasizes the principle of learning by doing, stating in the preface that "I believe that learning data management requires practice ..." (xxx). Therefore, he recommends reading the book while sitting in front of a computer and running Stata. He strives to use real-world data whenever feasible, with examples including the General Social Surveys for 2002, 2006, and 2016 and the National Survey of Youth, 1997. The book's target audience is students and practitioners in the behavioral and social sciences. Thus, the presentation of basic statistical modeling is supplemented with dis-

cussions of effect sizes and standardized coefficients. Various selection criteria, such as semipartial correlations, are discussed for model selection. He also covers various commands available for evaluating reliability and validity of measurements.

Acock's writing style is clear, concise, and straightforward. He uses simple language and structures his sentences and paragraphs logically, making it easy for readers to follow his arguments and ideas. Every chapter in the book starts with a table of contents and concludes with a review followed by a series of exercises. In addition, there are boxed tips spread throughout the text, providing useful discussions on various topics.

In this review, I summarize the contents of the book chapter by chapter, explicitly identifying what is new since the fourth edition. Major additions include a chapter on multilevel analysis for longitudinal and panel models as well as a chapter on item response theory (IRT), both introduced in the fifth edition. An extension to the `sem` command (structural equation modeling [SEM]) to fit regression models to account for missing values is added in the sixth edition. The revised sixth edition is updated for Stata 17, including updated discussion and images of Stata's interface and modern command syntax. In addition, examples include new features such as the updated `table` command and `collect` suite for creating and exporting customized tables as well as the option for creating graphs with transparency.[1]

2 Content

Chapter 1 begins by familiarizing the reader with the book's linguistic conventions. It then introduces the Stata interface and provides readers with a brief example of a Stata session. This includes loading a Stata dataset and generating basic summary statistics and graphs using the menus. In contrast to the fourth edition, the current edition includes a new subsection that delves into video aids for learning Stata.

Chapter 2 focuses on the creation of datasets. The chapter begins by addressing crucial issues related to data entry and verification. Using a sample questionnaire, it demonstrates how to establish a logical coding system. The chapter then covers the processes of entering data using Stata's Data Editor, managing variables using the Variables Manager, and saving a Stata dataset. Compared with the fourth edition, where only Excel files were discussed, the boxed tip at the end of the chapter now covers working with SAS and SPSS files.

Chapter 3 centers on dataset preparation for analysis. Acock highlights the value of meticulous planning in this phase and provides a sample outline for a data management project. The chapter covers several data management concerns, such as coding missing values, labeling variables and values, renaming variables, recoding values, generating new variables, and consolidating a set of variables into one scale.

Chapter 4 covers Stata commands, do-files, and results. Acock explains the fundamental structure of Stata commands through simple examples and delves into the significance and utility of do-files, providing instructions on creating, using, and saving

1. Discussion and examples include new features added to Stata since Stata 15.

a do-file. The chapter also outlines do-file management and best practices, such as adding comments, handling lengthy commands, and creating separate do-files for various tasks. Additionally, it briefly discusses copying and pasting results from the Results window to a word processor and saving results in a log file.

Chapter 5 deals with descriptive statistics and graphs for one variable. It begins by examining various measures of central tendency and variability, followed by a discussion of Stata commands for generating suitable descriptive statistics for both categorical and quantitative variables. Acock demonstrates how to create standard distributional plots using the menus and provides a brief introduction to the Graph Editor. In contrast to the fourth edition, the chapter includes an addition that showcases the updated `table` command in Stata 17. Acock uses this command to illustrate how users can obtain a statistical summary of a variable. Furthermore, he shows how to export the generated table in various formats using the new `collect export` command.

Chapter 6 focuses on analyzing the relationship between two categorical variables through measures of association and graphical representation. This section covers various techniques such as cross-tabulations and hypothesis tests for both ordered and unordered categorical variables and explores concepts such as probabilities, odds, and odds ratios specifically for binary outcomes. Additionally, Acock demonstrates how to summarize a quantitative variable across levels of a categorical variable using both a bar chart and the `table` command.

Chapter 7 deals with tests for one or two means. It covers topics such as randomization and random sampling and explains the concept of p-values. The discussion of means includes both independent and dependent t tests. The chapter also goes into detail on different measures of effect size and tests for unequal variance. Additionally, there is a section dedicated to power analysis that explains the process and assumptions underlying sample-size calculations. Finally, the chapter explores common nonparametric tests and provides examples. The boxed tip on effect size includes more comprehensive information. While the fourth edition covered only R^2 and Cohen's d, the current edition expands on this by including additional discussion on Hedges's g statistic and the point-biserial r. Furthermore, Acock provides guidance on how to use either the `table` command or a histogram with a normal density line to verify the normality assumption of the t test. The chapter concludes with a new subsection that provides a link to a video tutorial relevant to the topics discussed in the chapter.

Chapter 8 covers bivariate correlation and simple linear regression (that is, linear regression with one regressor). Acock demonstrates how one can create a basic descriptive scatterplot and add a regression line using the `twoway graph` dialog box. The chapter covers various aspects of correlation and rank correlation, including obtaining casewise and pairwise correlation coefficients, p-values, and adjusting for multiple comparisons. Furthermore, the chapter provides a brief overview of regression and explains how to use the `regress` dialog box to fit a simple linear regression model and interpret its output. The current edition features a new subsection at the end of the chapter on power analysis with correlation that explains how to use the `power onecorrelation` command to calculate sample size, power, or target correlation for a one-sample correlation test.

In chapter 9, analysis of variance (ANOVA) and its underlying logic and assumptions are introduced, followed by a simple hypothetical example that the reader is guided through. Acock takes care to explain the Stata output in detail and discusses the various options available for adjusting for multiple comparisons. Additionally, the chapter delves into alternative methods such as analysis of covariance and two-way ANOVA, as well as repeated-measures designs and intraclass correlation for measuring agreement. Readers are also introduced to Stata's official commands for obtaining adjusted means and graphs of predictive margins, that is, `margins` and `marginsplot`. Furthermore, the chapter covers a range of valuable commands for exploring and summarizing data in tables and graphs. Lastly, the suite of Stata `power` commands is introduced once again, with examples of power analysis for one-way, two-way, and repeated-measures ANOVA.

Chapter 10 deals with multiple regression. It starts with an introduction to its uses and a demonstration of how to fit such a model in Stata. The chapter then carefully explains and interprets the regression output, with a focus on model coefficients. Standardized beta weights, semipartial correlation, and the increment in R^2 are also discussed. In addition, the chapter covers a wide range of topics, including postestimation commands for checking model assumptions, multicollinearity issues, variance inflation factor calculation, weighted data handling, factor variables, interactions, and nonlinear associations testing. The chapter also showcases the use of the `margins` and `marginsplot` commands to obtain and plot adjusted predictions. Finally, the chapter concludes with a section on power analysis in multiple regression. Overall, this chapter offers a comprehensive and detailed exploration of multiple regression techniques in Stata.

Chapter 11 covers logistic regression. The chapter starts by discussing the unsuitability of ordinary least-squares regression for analyzing binary outcomes. Subsequently, the chapter introduces logistic regression and covers topics like odds, odds ratios, and the logit transformation. One of the boxed tips in the chapter explains the difference between the odds ratio and relative risk. Acock demonstrates how to fit a multivariable logistic model using the `logistic` dialog box and compares the output for the `logit` and `logistic` commands while carefully explaining the interpretation of coefficients. The chapter covers hypothesis testing for both single and multiple coefficients using likelihood-ratio and Wald tests, as well as a brief section on nested or hierarchical regression. The interpretation of model coefficients is further discussed, and examples of using the `margins` command to examine the effects of predictors are provided. The chapter also includes a section on power analysis in the context of logistic regression. In contrast to the fourth edition, the current edition features an additional subsection at the end of the chapter that provides links to tutorials related to logistic regression. These resources go beyond covering logistic regression for binary outcomes and additionally cover ordinal, count, and fractional outcomes.

In chapter 12, various topics related to measurement, reliability, and validity are discussed. The chapter emphasizes the importance of high-quality measurements in statistical analysis and highlights the issues that can arise because of poor measurements. Acock provides guidance on constructing a scale and illustrates how to calculate a mean score from a set of items. The chapter also delves into reliability, covering

correlation, intraclass correlation, alpha reliability, and measures of agreement such as kappa. Additionally, the concept of validity is explored, including expert-judgment, criterion-related, and construct validity. Factor analysis is discussed in depth, with Acock carefully explaining the terminology and techniques involved. The chapter concludes by presenting a detailed demonstration of performing a principal-components factor analysis.

Chapter 13 covers SEM and generalized SEM. This chapter is essentially what was the fourteenth chapter in the fourth edition with some extensions. Acock begins by using the SEM Builder to fit a basic multiple regression model, followed by a newly added subsection dealing with SEM and working with missing values. He discusses the `method(mlmv)` option of the `sem` command, which implements full information maximum likelihood, a simple way of handling missing values when the assumptions of missing at random and multivariate normality are reasonable. Furthermore, Acock demonstrates how one can add auxiliary variables to help justify the assumption of missing at random. The chapter then shows readers a much quicker way to fit the multiple regression model by using the regression tool in the SEM Builder and the `sem` command syntax. The next section in the chapter discusses generalized SEM, illustrating how to use the SEM Builder to fit a logistic regression model. The chapter explains how to obtain odds ratios for a 1-standard-deviation change and how to add these to the SEM diagram. Finally, the chapter concludes with a section that briefly extends these concepts to performing path analysis and another section that discusses causal models, mediation, and direct and indirect effects.

Chapter 14 deals with working with missing values and the technique of multiple imputation. This chapter corresponds with chapter 13 in the fourth edition. The chapter starts off with a new added section that introduces multiple imputation as an alternative to full information maximum likelihood, discussed in the previous chapter. It then proceeds with a discussion on what variables need to be included when doing imputations. Acock describes the nature of the problem of missing values and enumerates various reasons why missing values may occur in a study. The chapter cautions against using incomplete cases or ad hoc imputation methods such as imputing the mean value because this can lead to biased estimates and reduced power. To address this problem, the chapter walks the reader through a detailed example of performing multiple imputation using a multivariate-normal regression approach. Additionally, the chapter covers how to handle imputed values that are impossible, as well as the imputation of squared terms and interactions.

Chapter 15 was a new addition to the fifth edition and provides an introduction to multilevel analysis in Stata. Acock begins by describing data for groups of individuals and panel data, which are the two types of data appropriate for multilevel models. He also discusses the limitations of using linear regression for such data and introduces fixed-effects regression models and random-effects regression models, as well as the `xtreg` command for fitting these models. Furthermore, Acock provides examples of questions that can be asked of multilevel data and outlines how to handle panel data, including reshaping the data and visualizing them using the `twoway` command. He introduces the random-intercept model, which is appropriate under the assumption that individuals

have different intercepts, and shows how to fit this model using the `mixed` command. He also demonstrates how to visualize the model using `margins` and `marginsplot`. Next Acock considers the model with a quadratic term added on the right-hand side and the model where time is treated as a categorical variable. He then moves on to describe the random-coefficients model, which allows both slopes and intercepts to vary across individuals. He illustrates how to fit this model, again using the `mixed` command. Acock also shows how to include both time-varying and time-invariant covariates to explain the individual slopes in these models.

Chapter 16, the final chapter, was also a new addition to the fifth edition and covers IRT. Acock begins by contrasting IRT measures of variables with summated scales discussed in Chapter 12. He then provides an overview of three IRT models for binary items: the one-parameter logistic model, the two-parameter logistic (2PL) model, and the three-parameter logistic model. Acock then demonstrates how to fit both the one-parameter logistic and the 2PL models using Stata's `irt 1pl` and `irt 2pl` commands, respectively. He also illustrates and discusses various postestimation tools available for these fitted models. Furthermore, Acock considers an extension of the 2PL model to ordered categorical items, known as the graded response model, and shows how to fit this model using the `irt grm` command. He discusses the reliability of the fitted IRT model and demonstrates how to replicate the results of the commands using Stata's menu system. Finally, Acock discusses extensions of IRT, including the partial credit model and the hybrid model.

The book ends with a brief appendix that directs readers to additional resources for learning how to use Stata.

3 Assessment

As evidenced by the success of previous editions, Acock's book has been well received by Stata users. Therefore, rather than highlighting positive aspects of the book that have been covered in previous reviews and partly in the introductory section of this article, I will focus on discussing the new additions since the fourth edition and offering suggestions for improvement.

I was pleased to see that Acock finally wrote a chapter on multilevel analysis, which was long overdue. In the behavioral and social sciences, many studies involve repeated observations of the same individuals over time, or panel data. For example, experimental studies in economics often involve repeated observations because the investigator is interested in observing whether there are learning effects or whether there is convergence to some predicted or behavioral equilibrium. Although previous editions covered complex topics like multiple imputation, it was surprising to find that the analysis of panel data—such an essential topic—was missing.

Having said that, I am not completely satisfied with the presentation. While Acock was thorough in presenting the assumptions of multiple regression and walking the reader through different diagnostics, the presentation of panel-data models lacks the

same level of care and attention to detail. When presenting the fixed-effects model and before introducing the random-effects model, Acock rightly points out that if the time-invariant individual effects are uncorrelated with the right-hand side variables, then the random-effects model gives consistent estimates and is more efficient than the fixed-effects model (p. 466). This raises the question of how to verify this assumption.

As an instructor in the behavioral and social sciences, I often encounter the narrative that the choice between fixed effects and random effects or between estimators depends on the theory and the problem. Thus, if one is interested in within-panel variation over time, then fixed effects is the way to go. On the other hand, if the theory is about cross-panel differences, then one should not use fixed effects. I believe that mainstream textbooks, such as Acock's, have a role in dispelling this narrative. As Jeff Wooldridge puts it, "I don't think of it as a problem of 'variation' [...]. If you want to control for systematic, unmeasured differences across units (individuals, firms, schools, and so on) then [fixed effects] is preferred. If the variables of interest don't vary enough over time to identify the effects then we might need a new problem or a new data set." He further notes that "How can any theory reliabl[y] conclude that unobserved heterogeneity is uncorrelated with the observed covariates? How could I ever be sure that, say, managerial talent is unrelated to firm inputs? The only theory that implies [pooled OLS] or [random effects] is suitable is if we have a randomized intervention—still quite rare in the social sciences."[2]

Therefore, my recommendation is that Acock assert this point more forcefully and add a discussion on the Hausman test (implemented by the `hausman` command) and the test of overidentifying restrictions for panel data (implemented by the `xtoverid` command [Schaffer and Stillman 2006] from the Statistical Software Components Archive), which are standard tests used to verify the appropriateness of random effects. `xtoverid`, unlike `hausman`, reports a test statistic that is robust to arbitrary heteroskedasticity and within-group correlation if the `cluster` option was used by the original estimation.

The subsection dealing with SEM and working with missing values is a welcome addition. In my experience, I find that the theory of multiple imputation appears difficult when first explained to students. Therefore, preceding it with a description of full information maximum likelihood provides a simple and more intuitive way of introducing the topic of missing values and the underlying assumptions. Additionally, the chapter on IRT nicely complements the chapter on measurement, reliability, and validity.

Furthermore, Acock effectively highlights the updated `table` command and `collect` suite in Stata 17, demonstrating how customized tables can be created and exported using official Stata commands, which was previously left primarily to community-contributed commands. However, graphics is an area where community-contributed commands could be given more attention. It would be beneficial to showcase popular commands that can better visualize data than official commands in some cases, even if it is done through boxed tips. For instance, `tabplot` (Cox 2016) combines a table and a bar chart

2. https://www.statalist.org/forums/forum/general-stata-discussion/general/1519558-ols-vs-panel-regression.

to efficiently represent contingency tables for up to three categorical variables. Similarly, `coefplot` (Jann 2014) provides a more flexible alternative to the `marginsplot` command and can be used to plot regression coefficients. Including such illustrations would be valuable.

4 Acknowledgments

I thank the *Stata Journal* editors for extending an invitation to review Acock's book and Jochen Jungeilges, the aforementioned instructor, who introduced me to this book.

5 References

Acock, A. C. 2023. *A Gentle Introduction to Stata*. Rev. 6th ed. College Station, TX: Stata Press.

Collier, T. 2015. Review of Alan Acock's A Gentle Introduction to Stata, Fourth Edition. *Stata Journal* 15: 588–593. https://doi.org/10.1177/1536867X1501500216.

Cox, N. J. 2016. Speaking Stata: Multiple bar charts in table form. *Stata Journal* 16: 491–510. https://doi.org/10.1177/1536867X1601600214.

Jann, B. 2014. Plotting regression coefficients and other estimates. *Stata Journal* 14: 708–737. https://doi.org/10.1177/1536867X1401400402.

Mulcahy, M. 2006. Review of A Gentle Introduction to Stata by Acock. *Stata Journal* 6: 420–424. https://doi.org/10.1177/1536867X0600600310.

Schaffer, M., and S. Stillman. 2006. xtoverid: Stata module to calculate tests of overidentifying restrictions after xtreg, xtivreg, xtivreg2, and xthtaylor. Statistical Software Components S456779, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s456779.html.

About the author

Andrew Musau is an associate professor of economics at Molde University College. He obtained his PhD in economics from the University of Trento. His research interests are in behavioral and experimental economics, energy economics, and macroeconomics.