# The Stata Journal



The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

**Subscriptions** are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

**Subscription rates** listed below include both a printed and an electronic copy unless otherwise mentioned.

| U.S. and Canada | | Elsewhere | |
|---|---|---|---|
| **Printed & electronic** | | **Printed & electronic** | |
| 1-year subscription | $115 | 1-year subscription | $145 |
| 2-year subscription | $210 | 2-year subscription | $270 |
| 3-year subscription | $285 | 3-year subscription | $375 |
| 1-year student subscription | $ 85 | 1-year student subscription | $115 |
| 1-year institutional subscription | $345 | 1-year institutional subscription | $375 |
| 2-year institutional subscription | $625 | 2-year institutional subscription | $685 |
| 3-year institutional subscription | $875 | 3-year institutional subscription | $965 |
| **Electronic only** | | **Electronic only** | |
| 1-year subscription | $ 85 | 1-year subscription | $ 85 |
| 2-year subscription | $155 | 2-year subscription | $155 |
| 3-year subscription | $215 | 3-year subscription | $215 |
| 1-year student subscription | $ 55 | 1-year student subscription | $ 55 |

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.





The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, STATA, Stata Press, Mata, MATA, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2015)
**15**, Number 2, pp. 588–593

# Review of Alan Acock's A Gentle Introduction to Stata, Fourth Edition

Tim Collier
London School of Hygiene and Tropical Medicine
London, UK
Timothy.Collier@lshtm.ac.uk

**Abstract.** In this article, I review *A Gentle Introduction to Stata, Fourth Edition*, by Alan Acock (2014 [Stata Press]).

**Keywords:** gn0065, book review, introduction to Stata, data management, statistical analysis

## 1 Introduction

In this article, I review the fourth edition of Acock's *A Gentle Introduction to Stata* (2014), which has been updated to include features in Stata 13. It includes one completely new chapter on the `sem` and `gsem` commands, and many of the existing chapters have new material.

The book is principally aimed at social science students who need to perform statistical analyses but who have no prior experience with Stata or any other statistical software. The book's goal is to move such a reader from complete beginner status to competent Stata user. Acock's style is conversational and encouraging, and he enthusiastically asserts that Stata is "easy to learn and a joy to use" (p. 4).

The book is intended to be read sitting in front of a computer while running Stata and using the many example datasets that are freely available to be downloaded from the book's webpage. These datasets enable the reader to reproduce and further explore the results included in the book.

Throughout the book, Acock illustrates the methods being taught by using real data from various social science fields. The examples include data extracted from the General Social Surveys for 2002 and 2006 and the National Survey of Youth 1997.

The main focus throughout the book is on using Stata via the graphical user interface or dialog boxes, although the construction of Stata's command syntax is briefly explained and there is a short section on using do-files.

The book progresses logically from getting started with Stata to creating a dataset, checking data for problems, and preparing a dataset for analysis. It then explains descriptive statistics and basic hypothesis tests, analysis of variance, multiple linear regression, and logistic regression. There are also chapters on measurement, reliability and validity, multiple imputation, and then the new chapter on the `sem` and `gsem` commands. Acock has also added new sections to some existing chapters; for example,

 

there are new sections on power analysis with analysis of variance (ANOVA) and on nonlinear relations in multiple regression.

Each chapter begins with a contents list and ends with a review followed by a set of exercises. Interspersed throughout the book are 51 boxed tips that include helpful discussions on many topics such as searching for help or using Stata as a calculator, performing data management problems, or understanding statistical topics like effect size, statistical and substantive significance, and the concept of a strong kappa.

The book also contains screenshots of dialog boxes and Stata output (for example, Stata graphs and results), which help readers as they follow along.

## 2 Contents

Chapter 1 starts by introducing the book and explaining the conventions used throughout. Acock suggests that the book will probably be most helpful when read in front of a computer with Stata running. The book introduces the Stata interface, and it offers readers hands-on experience by providing a short example Stata session that includes loading a Stata dataset and producing some basic summary statistics and graphs using the menus.

Chapter 2 is concerned with creating datasets. It starts with discussing important issues surrounding data entry and verification and, using an example questionnaire, discusses how to develop a sensible coding system. It then discusses entering data using Stata's Data Editor, managing the variables using the Variables Manager, and saving a Stata dataset.

Chapter 3 focuses on preparing a dataset for analysis. Acock emphasizes the importance of careful planning in this process and gives an example outline for a data management project. The chapter deals with data management issues such as coding missing values, labeling variables and values, renaming variables, recoding values, generating new variables, and creating a single scale from a set of variables.

Chapter 4 deals with Stata commands, do-files, and results. Throughout, the book introduces Stata commands using the menus and dialog boxes, with a focus on using Stata interactively. However, chapter 4 explains the basic structure of Stata commands with simple examples. The chapter discusses the importance and usefulness of do-files and provides instructions on how to create, use, and save a do-file. It also discusses management and good practice for do-files including adding comments, dealing with long commands, and having separate do-files for different tasks. Acock then touches on copying and pasting results from the Results window into a word processor and saving results in a log file.

Chapters 5 to 14 cover various statistical topics from simple descriptive statistics and bivariate association through linear and logistic regression, to multiple imputation and structural equation models. That does not mean that data management issues have been left behind entirely; many of the statistical chapters include contextual data-

processing tasks. Similarly, while chapters 5 and 6 focus on simple descriptive statistics and measures of association, the later chapters focus more on advanced statistical topics with basic exploratory analyses and graphics.

Chapter 5 covers descriptive statistics and graphs for a single variable. It starts by discussing different measures of location and dispersion before proceeding to look at Stata commands for obtaining appropriate descriptive statistics for categorical and quantitative variables. The chapter covers user-written (for example, `fre`) as well as official Stata commands and shows how to install such commands. Using the menus, Acock walks the reader through various standard distributional plots and very briefly introduces the Graph Editor.

Chapter 6 is concerned with measures of association and graphs for two categorical variables. The chapter covers cross-tabulations and hypothesis tests for both ordered and unordered categorical variables, and it discusses probabilities, odds, and odds ratios for binary dependent variables. The chapter also introduces one of Stata's immediate commands (`tabi`) in the context of cross-tabulations and provides readers with the first of several encounters in the book with power analysis using a user-written command. Slightly out of context with this chapter, the book demonstrates how to summarize a quantitative variable over levels of a categorical variable in a bar chart and by using the `table` command.

Chapter 7 covers commonly used hypothesis tests of proportions and means for one or two samples. The chapter discusses the differences between and implications of randomization and random sampling and introduces the concept of hypothesis testing and $p$-values. The comparisons of means covers both independent and dependent (or paired) $t$ tests. This chapter also discusses different measures of effect size and tests for unequal variance. It also includes quite a long section on power analysis that introduces Stata's suite of `power` commands, and it includes a good discussion of the process and assumptions underlying sample-size calculations. The chapter concludes with some examples and discussion of common nonparametric tests.

Chapter 8 introduces bivariate correlation and simple linear regression (that is, linear regression with one independent variable) and related commands. Using the `twoway graph` dialog box, the book demonstrates how to produce a simple descriptive scatterplot and then how to overlay a regression line. Acock also demonstrates `binscatter`, a very neat user-written graph command. Many pages are dedicated to discussing correlation and rank correlation and to explaining how to obtain casewise and pairwise correlation coefficients and how to obtain $p$-values and adjust for multiple comparisons. This chapter briefly introduces the topic of regression and explains how to fit a simple linear regression model using the `regress` dialog box and how to interpret the output.

Chapters 9 and 10 cover ANOVA and multiple linear regression, respectively. They are the longest chapters in the book, together making up 25% of the whole book.

Chapter 9 begins by introducing ANOVA and its logic and assumptions before taking the reader through a simple hypothetical example. Acock carefully explains the Stata output and discusses the various options for making adjustments for multiple

comparisons. The chapter also covers a nonparametric alternative to ANOVA, analysis of covariance, two-way ANOVA, repeated-measures designs, and intraclass correlation for measuring agreement. The chapter provides readers with their first exposure to Stata's excellent `margins` and `marginsplot` commands for obtaining adjusted means and graphs of predictive margins. A variety of useful commands for exploring and summarizing data in tables and graphs are also covered in the chapter. The chapter further introduces the suite of Stata `power` commands, first encountered in chapter 7, with examples of power analysis for oneway, twoway, and repeated ANOVA.

Chapter 10 begins by introducing multiple regression and discussing its uses before showing how to fit such a model in Stata. The chapter then walks us slowly and thoroughly through the regression output, focusing on explaining and interpreting the model coefficients. This chapter also discusses standardized beta weights, semipartial correlation, and the increment in $R^2$. The chapter covers a lot of ground including postestimation commands for checking model assumptions, issues surrounding multicollinearity, how to calculate the variance inflation factor, how to work with weighted data, and how to include and test factor variables, interactions, and nonlinear associations. It also demonstrates once again how to use the `margins` and `marginsplot` commands to obtain and plot adjusted predictions. The chapter concludes with a section on power analysis in multiple regression using a user-written command called `powerreg`.

Chapter 11 covers analysis of a binary outcome using logistic regression. It starts with an example of a study with a binary outcome and explains why ordinary least-squares regression is inappropriate in such a setting. It then follows by introducing logistic regression and discussing odds, odds ratios, and the logit transformation; one of the boxed tips in this chapter discusses the odds ratio versus relative risk. Acock then shows how to fit a multivariable logistic model using the `logistic` dialog box. Here the output for the `logit` and `logistic` commands is compared and the interpretation of the coefficients is carefully discussed. A section follows on hypothesis testing for both single and multiple coefficients using likelihood-ratio and Wald tests, as does a short section on nested or hierarchical regression. The chapter then discusses the interpretation of model coefficients in more detail and gives some examples of uses of the `margins` command to examine the effects of predictors. The chapter ends with a section on power analysis for logistic regression using another user-written command, `powerlog`.

Chapter 12 covers issues regarding measurement, reliability, and validity. It begins by briefly discussing the importance of the quality of measurements in statistical analysis and the problems caused by poor measurements. Acock then discusses methods for constructing a scale and, as an example, shows how to generate a mean score from a set of items. The chapter then provides a fairly lengthy section on reliability (covering correlation and intraclass correlation, alpha reliability, and measures of agreement including kappa) and validity (covering expert-judgment, criterion-related, and construct validity). It then discusses factor analysis in detail, and Acock carefully explains the accompanying special vocabulary and the different techniques involved. The book follows this with a detailed example of carrying out a principal-components factor analysis.

Chapter 13 covers missing data and multiple imputation. The chapter starts by discussing the nature of the problem of missing values and lists different reasons why missing values may occur in a study. Acock then explains that analyses that involve using complete cases only, or that use ad hoc imputation methods (for example, imputing the mean value), result in reduced power and potentially biased estimates. The chapter then follows with discussing the assumptions of multiple imputation—that is, the different mechanisms for missingness—and some practical guidance on what variables to include when doing imputations. The book then walks the reader carefully through a detailed example of performing multiple imputation using a multivariate-normal regression approach. This includes a helpful section on preliminary analysis using the `misstable` command. The chapter also discusses what to do when impossible values are imputed and discusses the imputation of squared terms and interactions.

Chapter 14 is a completely new chapter to this edition, and it introduces the `sem` (structural equation modeling) and `gsem` (generalized structural equation modeling) commands. Acock uses the SEM Builder to fit a basic multiple regression model and then demonstrates how to improve the resulting output. The chapter then shows readers a much quicker way to build the same model by using the regression tool in the SEM Builder and by using the `sem` command syntax. The next section in the chapter discusses generalized structural equation modeling and shows how to use the SEM Builder to fit a logistic regression model. The chapter also explains how to obtain odds ratios for a 1-standard-deviation change and how to add these to the SEM diagram. The chapter concludes with a section that briefly extends these concepts to performing path analysis and a section that discusses causal models, mediation, and direct and indirect effects.

The book concludes with a short appendix that points the reader to further resources for learning how to use Stata.

# 3 Conclusion

Overall, this book does exactly what it says on the cover. It gives the reader a gentle introduction to Stata. Acock recommends reading the book while sitting in front of a computer running Stata; that is what I did for this review, and I agree that readers will get the most out of the book when using this method. The book could be used by a complete novice and worked through page by page as a do-it-yourself Stata course, or it could be used only as required.

Although some of the statistical methods covered (for example, principal-component factor analysis and structural equation modeling [SEM]) are certainly not simple and would not be covered in an introductory statistics course, Acock does introduce these advanced methods gently. For example, he introduces SEM by using it to fit a basic regression model. As might be expected with an introductory textbook that covers various statistical methods, some of the methods (particularly the more advanced ones) are only lightly discussed. Acock acknowledges, for example, that he has only scratched the surface of Stata's multiple-imputation capabilities and has given us just a taste of what the `sem` command can do.

The book is clearly meant for and will be found most useful by social science students. I am sure students from other disciplines would find the book helpful, but they might struggle with some of the language and concepts. For example, with my background in medical statistics, I found it difficult to imagine being in the public relations department for a small liberal arts college (p. 149).

I liked how many of the examples and exercises in the book involve performing data management tasks and how Acock begins by using exploratory analyses, often using both tables and graphs, before discussing the detailed statistical modeling aspects of each chapter. The exercises at the end of each chapter are also helpful.

Obviously you cannot cover everything in a textbook this size, but there were a few omissions that surprised me. I was surprised that there was no coverage of combining datasets using either `append` or `merge`. Although the example datasets were often large and interesting, they were all single datasets. In my experience, most analyses I carry out involve having to combine several datasets to get to an analysis-ready dataset. The goal of this book is to enable students to carry out analyses of their own data, and I wonder whether this omission might cause some students to fall at the first hurdle. I also expected to see some coverage of error messages, such as some discussion of the common error messages and information on how to interpret them.

In conclusion, I think this is a well-written introductory textbook that will be very helpful, particularly to social science students wanting to learn to use Stata. I have often recommended earlier editions to students who have attended courses that I have taught, and I will continue to do so with this edition.

# 4 Reference

Acock, A. C. 2014. *A Gentle Introduction to Stata*. 4th ed. College Station, TX: Stata Press.

**About the author**

Tim Collier is a senior lecturer in the Department of Medical Statistics at the London School of Hygiene and Tropical Medicine. His research is mainly focused on cardiovascular clinical trials and epidemiology. Over the past 15 years, he has enjoyed teaching Stata courses for medical statisticians and health researchers including the UK Stata Summer School.