Psychological Statistics and Psychometrics Using Stata

SCOTT A. BALDWIN
Brigham Young University

A Stata Press Publication
StataCorp LLC
College Station, Texas

Published by Stata Press, 4905 Lakeway Drive, College Station, Texas 77845
Typeset in LaTeX 2ε
Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

Print ISBN-10: 1-59718-303-2
Print ISBN-13: 978-1-59718-303-1
ePub ISBN-10: 1-59718-304-0
ePub ISBN-13: 978-1-59718-304-8
Mobi ISBN-10: 1-59718-305-9
Mobi ISBN-13: 978-1-59718-305-5

Library of Congress Control Number: 2019935130

To Autumn

Contents

(Pages omitted)

1 Introduction

Statistics is a major part of research in many fields, often serving as the primary method for establishing whether results support hypotheses. Indeed, it is unusual for a quantitative study in psychology to not use statistical analysis. Consequently, it's essential for psychologists to be competent users of statistics, and that is the primary aim of this book.

That is a lofty goal. Statistics is a huge, technical field. Psychologists who spend their careers studying statistics only master small portions, and most psychologists study psychological content, not statistics. Furthermore, studying a single book will not make you a competent user of statistics. Competence comes from trial and error and from applying statistical methods and ideas to your own research. Competence comes from years of building your skills. Competence comes from collaborating with more competent people and learning from them. My hope is that this book serves as a step toward competence for those starting their training and as a useful reference or development tool for those already using statistics.

1.1 Structure of the book

This book covers some foundational topics in psychological statistics and psychometrics. Topics include t tests, analysis of variance (ANOVA), regression, power analysis, multilevel models, structural equation modeling, and factor analysis. I do not provide an exhaustive treatment of each topic because many books already do that. Instead, I provide an introduction to key concepts regarding how to use these statistical models to answer research questions.

Competent statistical analysis requires the use of statistical software. Consequently, I weave the statistical content with Stata code that illustrates how to fit models and make sense of the output. Most of the code examples illustrate how to use Stata commands, such as `regress` (for regression) or `sem` (for structural equation modeling). I also illustrate how to use some of Stata's graphical commands, such as `histogram` and `twoway scatter` for scatterplots, and data management commands, such as `generate` and `egen` (to create new variables). I show how to program Stata to run simulations to help you learn a concept or as a method for understanding your models.

The book is divided into three parts. Part I provides an introduction to Stata, including the interface, loading data, do-files and Stata syntax, descriptive statistics, graphics, and help and documentation. Part II discusses regression, ANOVA, multilevel

models, and statistical power. I begin with regression because it provides the foundation for the discussion of ANOVA, multilevel models, and power. Part III covers psychometrics, including a discussion of the issues of reliability and validity. In these chapters, I discuss these concepts from the perspective of factor analysis, because I believe this provides a coherent framework for thinking about these measurement concepts. Looking at these concepts from the perspective of factor analysis will also help you make connections between the psychometric concepts of reliability and validity and regression models.

I *strongly* recommend that you work through the Stata code in each chapter rather than just reading the code and seeing the results in the book. You will learn statistical concepts better by toying with code and seeing what happens, especially as you change the specification or the options. My students tell me that it is more effective to practice using Stata and to run models during class rather than just watch me use Stata or read lecture notes. Furthermore, as you get more comfortable with Stata, you will find it much easier to learn how to use other Stata commands because you will start to see the connections between the various commands.

1.2 Benefits of Stata

Statistical software is necessary, but why choose Stata? Some treat statistical software kind of like a sports team, showing undying devotion to the software and viewing criticisms of the software as personal attacks. Others are pragmatic users, simply using what is available to them or what seems useful. Some want only open source software; others want a company to back the development and certification of the software. Many use a specific software because that's what their adviser used or what was taught in their program. I started using Stata because I was a longtime Mac user. As I was finishing my graduate degree in 2003, SAS did not support Mac at all, and SPSS's Mac software was buggy and slow. Stata's Mac support was excellent, and the software did most of what I needed. I had been introduced to Stata as an undergraduate and was happy to return. I have never looked back. Stata has remained my primary statistical package.[1]

Although my reasons for choosing Stata had little to do with statistics itself, I believe Stata is an excellent choice for five reasons:

1. **Consistent syntax**. Learning Stata, like learning any programming language, is challenging. However, learning Stata is manageable because Stata's syntax is expressive and consistent. It is expressive because it is easy to understand once you have the basics down. It is consistent because command syntax does not fundamentally change from one command to another. Thus, once you get the basic structure down, it is easy to learn how to use new commands: the structure will be similar to commands you are already familiar with.

1. That's not to say that I do not use other software. Sometimes, collaboration requires that I do. Sometimes, I need analysis routines not available in Stata. And sometimes, I just like to learn new stuff, so I try software that is available to me.

2. **Aids the replicability of analyses**. Statistical analyses should be reproducible (Long 2009), which means that if I asked you to run an analysis a second time (or third, fourth, and so on), you will get the same answer. Likewise, if I ran your code on my computer, I would get the same results. Stata includes the ability to specify the Stata version number for your analysis. For example, you may start your analysis with the command `version 14.2`, which tells Stata to run the analysis using the code that was present for Stata version 14.2. Therefore, if something changed between version 14.2 and 15.0, you will not get different answers.[2]
3. **Comprehensive documentation**. Stata's support documentation is comprehensive and ships as part of the software. Additionally, all documentation is freely available on the web. In my opinion, Stata documentation is second-to-none in the statistical software world. The documentation provides a readable explanation of each command, including options. Furthermore, the documentation includes worked examples and discussion of output to aid your learning. Finally, the documentation provides technical details regarding the mathematical and statistical underpinnings of commands. You can learn a lot by studying the documentation.
4. **Data management**. Most analyses require data management: cleaning data, generating variables, labeling variables and values, reshaping the data into a specific format, and so on. Stata's data management commands and utilities are excellent and make data preparation straightforward and replicable. Even when I need to use another software package for analysis, I nearly always prepare and manage the data in Stata.
5. **Graphics**. Graphics are an essential part of data analysis and are often superior to tables of numbers when it comes to communicating results (Cox 2004; Gelman, Pasarica, and Dodhia 2002; Gelman 2011). Stata includes comprehensive graphical tools to aid in exploring your data and interpreting the results of your models. The graphics are customizable and can be quite beautiful. Nearly all figures in this book were created with Stata. I include the code for creating these figures so that you can reproduce the graphs. Study this code. You will appreciate how flexible Stata graphs can be, and you will learn about Stata programming.

1.3 Scientific context

As I write this introduction, psychology, and science generally, has some problems. Pressure to publish and to ensure it is something exciting and novel combined with bad methodological practices means that a lot of research is not replicable (Ioannidis 2005, 2008, 2012). Indeed, many psychological studies simply do not replicate (Open Science Collaboration 2015), leading some to call the situation a "replication crisis"

2. Of course, if the Stata developers caught a bug in the code that results in a different answer in version 15 than 14.2, you may want the different answer. Nevertheless, by making the version number explicit, you can ensure that the results will only change when you expect them to.

(Pashler and Harris 2012). Psychology is not alone in this, with some pointing to problems in other disciplines, such as economics, biology, and medicine (Ioannidis 2005, 2013, 2014). Do these problems mean that many (or even most) theories are not worthwhile and the scientific literature cannot be trusted? I hope it is not that bad. Regardless of the answer to that question, I think the replication crisis does suggest we ought to step back and think about why we are facing these problems.

Some argue that a major reason for these problems is the incentives that influence scientists (Baldwin 2017; Ioannidis 2014; Nelson, Simmons, and Simonsohn 2012; Simmons, Nelson, and Simonsohn 2011). For example, in universities across the world, promotion and tenure depend upon publications—hence the phrase "publish or perish". Getting your first academic job often depends upon having many publications, including some in prestigious journals. Some research positions are "soft money" jobs, which means that salary and research support comes from grant money rather than the university itself. Getting grants requires that you are productive and that previous grants worked out. Sometimes we joke that to get a grant, we have to do all the research that the grant is proposing to prove that the research will work. When your salary and reputation depend upon getting papers published and the research turning out in a specific way, the incentive is to make sure the research works out as predicted.

Given these incentives, publishing becomes the end goal of research. That is, rather than publishing being the means to communicate scientific observation, publishing and adding lines to your vita become tantamount to science itself. The book *The Compleat Academic*[3] provides advice to researchers in the early stages of their careers on things like graduate school, applying for postdocs and jobs, submitting grants, and teaching classes. The advice to graduate students states:

> The information that we need to arrive at a short list of applicants is contained in the letters of recommendation and, primarily, in the academic vita. Wise graduate students, therefore, will start at day one of their first year in a PhD program to develop a strong vita. [...] Alter your perspective so that you derive your professional self-respect entirely from what is on that document. From the start of graduate school on, throughout what we hope will be a long and productive career, you *are* your vita. (Lord 2004, 10, emphasis in original)

Given such advice, combined with the incentives for getting and keeping a job (including securing your own salary!), it is not difficult to see why publishing became equivalent to doing science.

I learned statistics in this context, as did most researchers before me. Consequently, a number of problematic research and statistical practices evolved that ultimately helped publication rates but did not improve the quality of the science. For example, consider the use of p-values to evaluate statistical significance. There are many criticisms of

3. The dictionary on my computer defines the word "Compleat" as "archaic spelling of complete". Leave it to academics to take the simple word *complete* and make it snooty.

p-values in the scientific literature (for example, Meehl [1978]). Many of the criticisms are about how p-values are used, not so much about p-values themselves. McElreath (2016) says it well:

> This audience accepts that there is something vaguely wrong about typical statistical practice in the early 21st century, dominated as it is by p-values and a confusing menagerie of testing procedures. [...] The problem in my opinion is not so much p-values as the set of odd rituals that have evolved around them, in the wilds of the sciences, as well as the exclusion of so many other useful tools. (pp. xi–xii)

In short, p-values became the primary source of evidence that a result is publishable. Consequently, the goal of analysis becomes finding a p-value that is less than 0.05. If you complete a study and find null results, there's a good chance you will not even try to publish it.

Researcher flexibility, especially with respect to design, analysis, and reporting, means that finding significant effects probably required torturing the data. More advice from *The Compleat Academic*[4] explains this well:

> To compensate for this remoteness from our participants, let us at least become familiar with the record of their behavior: the data. Examine them from every angle. Analyze the sexes separately. Make up composite indexes. If a datum suggests a new hypothesis, try to find additional evidence for it elsewhere in the data. If you see dim traces of interesting patterns, try to reorganize the data to bring them into bolder relief. If there are participants you do not like, or trials, observers, or interviewers who gave you anomalous results, drop them (temporarily). Go on a fishing expedition for something—anything—interesting. (Bem 2004, 187)

"Interesting patterns" here usually means small p-values. Such flexibility in analysis is sometimes called researcher degrees of freedom or p-hacking (Simmons, Nelson, and Simonsohn 2011) or the garden of forking paths (Gelman and Loken 2014). Examples of such analyses include running multiple experiments and only reporting results from those with significant results. If main effects are not significant, test interactions; dropping observations when doing so takes the p-value from $p = 0.09$ to $p = 0.03$. Change a continuous variable to a categorical variable because the categorical variable produces significant results. Ignore problems with estimation and fit because a result is statistically significant. Fail to look at the raw data to see what a model (which is a reduction of the data) implies about the data because the analysis is statistically significant (Simmons, Nelson, and Simonsohn 2011). I think you get the idea.

4. Honestly, I cannot get over that name.

Fortunately, changes are in process. For example, the Open Science Framework (https://osf.io/) provides tools for researchers to register hypotheses before seeing the data and to create a website for hosting data and analysis files. I have seen job postings where an emphasis on improving rigor and replicability of science is a job qualification. Journals are accepting registered reports, wherein studies are reviewed prior to data collection and evaluated solely on the basis of the research question and quality of the design and proposed analyses (https://cos.io/rr/).

I hope this book contributes to the positive changes. My goal is to teach statistical concepts and software in a way that helps researchers a) address their research question transparently and openly, b) better understand their data, and c) better understand the models they use and what the models imply about their theory or research area. To be sure, I teach and use p-values—they can be useful. Stata includes many tools that supplement the information provided by p-values, and as a consequence, Stata can be used in a way that improves how statistics are used in psychology specifically and in science generally. So buckle up! We have a lot of great stuff to discuss.

(Pages omitted)

3.3 Bivariate regression

The mean was our best guess regarding the value of `new_attitude`. Another term meaning best guess is *expected value*. That is, in the long run, what do we expect the value of `attitude` to be? For a single variable, such as `attitude`, the expected value is the mean. Regression, on the other hand, allows us to determine an expected value for `attitude` that incorporates information from other variables. Our best guess about `attitude` will be based on variables such as the respondent's level of education and mental health symptoms.

Figure 3.2 is a scatterplot of `educ` and `attitude`.

```
. use http://www.stata-press.com/data/pspus/gss_attitude, clear
. twoway scatter attitude educ, jitter(2) scheme(lean2)
```

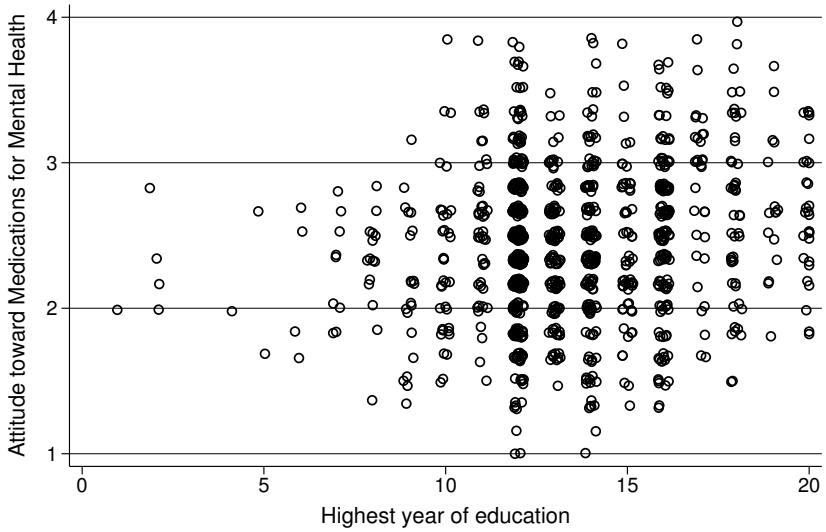

Figure 3.2. Relationship between `educ` and `attitude`

Based on figure 3.2, what do you think is the expected value for `attitude`? Does that expected value differ for someone who has 10 years of education compared with someone who has 20 years of education? That is, if we consider `educ`, does that affect what we expect `attitude` to be? If there is a relationship between education and attitude toward psychotropic medication, then expected values of `attitude` will vary as a function of `educ`. Regression provides information about the nature of that relationship in the form of a line, which is why regression is also called linear regression.

3.3.1 Lines

Let's go back in time and revisit junior high school algebra. You may recall the formula for a line:

$$y = mx + b$$

where x and y are variables, m is the slope, and b is the y intercept. We say that y is a function of two quantities: i) the product of m and x and ii) b. The slope describes changes in y with respect to changes in x (that is, "rise over run"), and the y intercept provides the value of y when $x = 0$. If we know m, x, and b, then we know y (see figure 3.3). Regression produces slope and intercept values that describe the relationship between an outcome and predictors.

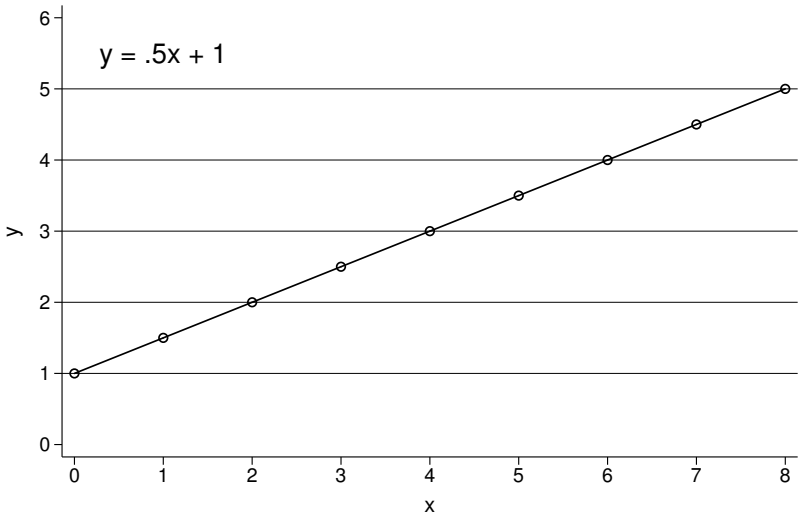

Figure 3.3. Illustration of $y = mx + b$

3.3.2 Regression equation

Compare figures 3.2 and 3.3. In figure 3.3, it is obvious what the line should be because the points all fall on a straight line. In figure 3.2, you could draw any number of lines through the data—a horizontal line, a vertical line, a line moving up from left to right, or a line moving down from left to right. All lines would provide some indication of the relationship between `attitude` and `educ`, and all lines will provide an intercept and slope. Which line is best? The best line is the line that provides the closest correspondence between the expected value of y and the actual value of y. In other words, the regression line will provide the best prediction of y given the x values across the entire dataset.

A regression equation for the relationship between `attitude` and `educ` is

$$\texttt{attitude}_i = \beta_0 + \beta_1 \texttt{educ}_i + \epsilon_i \tag{3.1}$$

This equation says that the observed value of `attitude` for person i is a function of an intercept (β_0), a slope (β_1) relating `educ` to `attitude`, and error in prediction (ϵ_i). Error is included in (3.1) because the observed value of `attitude` is not perfectly captured by β_0 and β_1 (that is, the data do fall on a straight line).

Commonly, the regression equation is written in terms of the expected value of y. For the example model, the equation is

$$E(\texttt{attitude}_i|\texttt{educ}_i) = \beta_0 + \beta_1 \texttt{educ}_i \tag{3.2}$$

which is read as, "The expected value of `attitude` given `educ` is equal to β_0 plus β_1 times `educ`." Alternatively, we can substitute $E(\texttt{attitude}_i|\texttt{educ}_i)$ with $\widehat{\texttt{attitude}}_i$:

$$\widehat{\texttt{attitude}}_i = \beta_0 + \beta_1 \texttt{educ}_i \tag{3.3}$$

The "hat" over $\texttt{attitude}_i$ denotes the predicted value of `attitude`. Equations (3.2) and (3.3) omit ϵ_i because these equations deal with the expected or predicted values of `attitude`—errors come into play when comparing the expected value to the actual value of y.

A general bivariate regression (one y and one x) equation is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{3.4}$$

Further, the general form of (3.2) and (3.3) is

$$E(y_i|x_i) = \beta_0 + \beta_1 x_i \tag{3.5}$$

and

$$\widehat{y}_i = \beta_0 + \beta_1 x_i \tag{3.6}$$

Population parameters versus sample-based estimates

In research, we collect data on samples to learn about populations. In regression, we estimate slopes and intercepts based on samples to learn about what the slopes and intercepts are in the broader population. Slopes and intercepts in the population are called population parameters. Parameters are typically symbolized using Greek letters, such as β in the case of regression coefficients [see (3.4)]. In this book, sample-based estimates of those equations are symbolized using lowercase, Roman letters, such as b in the case of regression coefficients. Thus, the population slope is β_1 and the sample-based estimate of the slope is b_1. Other examples include σ and s for the population and sample-based standard deviation, and ρ and r for the population and sample-based correlation.

3.3.3 Estimation

The slope is computed as

$$b_1 = r_{yx}\frac{s_y}{s_x} \tag{3.7}$$

where r_{yx} is the correlation between x and y, and s_y and s_x are the standard deviations for y and x, respectively. The intercept is computed as

$$b_0 = \overline{y} - b_1\overline{x} \tag{3.8}$$

Of course, we rarely—if ever—compute these quantities by hand. Stata's `regress` command does this for us. The syntax for `regress` is straightforward. Following the keyword `regress`, you type the dependent variable followed by the independent variables. As discussed in section 2.4.1, you can use `if` and `in` with `regress`.

`regress` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] [, *options*]

Estimating (3.3) with `regress` is done as follows:

```
. regress attitude educ
      Source |       SS           df       MS      Number of obs   =     1,006
-------------+----------------------------------   F(1, 1004)      =     33.02
       Model |  8.81344558         1  8.81344558   Prob > F        =    0.0000
    Residual |  267.958141     1,004  .266890579   R-squared       =    0.0318
-------------+----------------------------------   Adj R-squared   =    0.0309
       Total |  276.771587     1,005  .275394614   Root MSE        =    .51661

------------------------------------------------------------------------------
    attitude |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0335253    .005834     5.75   0.000     .0220771    .0449736
       _cons |   1.991708   .0804314    24.76   0.000     1.833876    2.149541
------------------------------------------------------------------------------
```

Focus on the bottom portion of the output for now. Stata calls the estimated intercept `_cons` (that is, constant). Thus, $b_0 = 1.99$. The slope for `educ` is $b_1 = 0.03$. We can now fill in the details of (3.3).

$$\widehat{\texttt{attitude}}_i = 1.99 + 0.03\texttt{educ}_i \tag{3.9}$$

We can make predictions based on the coefficients. For example, if we want to know the predicted `attitude` for someone who has 15 years of education, we simply plug 15 into (3.9).

$$2.44 = 1.99 + 0.03 \times 15 \tag{3.10}$$

The predicted `attitude` for someone who has 15 years of education is 2.44. Another way to say that is, "The expected value of `attitude` given that `educ` equals 15 is 2.44, or $E(\texttt{attitude}|\texttt{educ} = 15) = 2.44$."