

THE ECONOMICS COLLECTION

Philip J. Romero and Jeffrey A. Edwards, *Editors*

# A Beginner's Guide to Economic Research and Presentation

Jeffrey A. Edwards



[www.busessexpertpress.com](http://www.busessexpertpress.com)



**Balance Sheet**

Miles and Associates

56,880 (6%)

27,221 (3%)

# **The Economics Collection**

**Philip J. Romero and Jeffrey A. Edwards, *Editors***

# **A Beginner's Guide to Economic Research and Presentation**

**Jeffrey A. Edwards**



**[www.businessexpertpress.com](http://www.businessexpertpress.com)**

# to Economic Research and Presentation to Economic Research and Presentation

Jeffrey A. Edwards



*A Beginner's Guide to Economic Research and Presentation*

Copyright © Business Expert Press, 2013.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations, not to exceed 400 words, without the prior permission of the publisher.

First published in 2013 by

Business Expert Press, LLC

222 East 46th Street, New York, NY 10017 [www.businessexpertpress.com](http://www.businessexpertpress.com)

ISBN-13: 978-1-60649-832-3 (paperback) ISBN-13: 978-1-60649-833-0 (e-book)

Business Expert Press The Economics collection

Collection ISSN: 2163-761X (print) Collection ISSN: 2163-7628 (electronic)

Cover and interior design by Exeter Premedia Services Private Ltd. Chennai, India

First edition: 2013

10 9 8 7 6 5 4 3 2 1

Printed in the United States of America.

## **Abstract**

Conducting good research is critical to any student today. Writing good research papers is equally important—yet many students have not been given the proper tools to convey cogently the results of their research. *A Beginner's Guide to Economic Research and Presentation* is intended to address and redress this need.

This book is literally a step-by-step approach to the writing of an undergraduate or graduate level research paper in the field of economics. The primary audience for this book consists of those students who have not conducted research or written a research paper, or those students that are looking for ways of improving their research skills. Most books concerned with research writing are broadly applied. They approach the subject generally, which is to say that they don't lay out a *particular* path to conducting

research. Yet a specific path offering a specific focus to writing research is exactly what is needed for most students. This book provides that focus. For example, *A Beginner's Guide to Economic Research and Presentation* doesn't cover a dozen different search engines to perform a literature review; it specifies only EconLit. Nor is the student left to decide what scholarly publications are important ones to review; the book emphasizes only the use of journal impact factors found through RePEc to rank journal articles and their importance to the literature at large. Whereas other books provide an overview of how to present research, with only cursory suggestions and tips, *A Beginners Guide to Economic Research and Presentation* provides precise details on all aspects of research writing, including how many PowerPoint slides one should prepare for presentations and how much content should be on each slide. In short, unlike other books, this book provides a *specific* approach to conducting research, writing a paper, and presenting its material.

### **Keywords**

regression, regression models, inference, equation editing, table design, research topics, research questions, research hypotheses, literature reviews, data collection, formatting, drafting, presentation

*This book is dedicated to my wife Catherine.*

# Contents

<i>Introduction</i> .....	xi
Chapter 1 Chapter 2	
Chapter 3 Chapter 4	
Chapter 5	
Chapter 6	
Chapter 7	
<a href="#">Chapter 8</a>	
<a href="#">Basic Regression Analysis and Inference</a> .....	1
<a href="#">More Sophisticated Regression Analysis and Inference</a> .....	13
<a href="#">Basic Equation Editing and Table Design</a> .....	31
<a href="#">Choosing a Research Topic, Question, and Hypothesis</a> .....	39
<a href="#">Literature Reviews</a> .....	45
<a href="#">Data Collection and Formatting</a> .....	53
<a href="#">Drafting and Refining the Paper</a> .....	61
<a href="#">Pointers on Presenting to a Live Audience</a> .....	71
<a href="#">Conclusion</a> .....	77
<a href="#">Notes</a> .....	79
<a href="#">Index</a> .....	81

# Introduction

This book is an introduction to writing a research paper in the field of economics. The primary audience for this book consists of students who have either not conducted research and/or written a research paper, or only broached the activity in a cursory manner. Such an audience might include undergraduate business majors in general, economics majors in particular, or graduate students studying in fields related to economics. And while this last group will probably find the first chapter on econometrics too rudimentary, the remainder of the book will be of great value even to that population of individuals.

For the instructor, the book is organized in a way that I have found yields the most understandability among the students when faced with performing research and writing a paper for the first time; but it is not written in the way other similar books are. Specifically, most books in this area are written as broadly applied guides to writing research papers. By broadly applied I mean that they don't lay out a *particular* path to conducting research. This is a problem simply because it is exactly that path that is needed for most students; this book provides that focus.

As mentioned in the abstract, this book doesn't cover a plethora of search engines to perform a literature review; it specifies only the primary search engine in economics, EconLit. It doesn't leave it to the student to decide what scholarly publications are important ones to review; it emphasizes only the use of journal impact factors found through RePEc to rank journal articles and their importance to the literature at large. This book doesn't just provide an overview of how to present research with some cursory suggestions and tips; it provides precise details on how many PowerPoint slides one should have and exactly how much content should be on those slides. In other words, unlike other books, this book provides a *specific* approach to conducting research, writing a paper, and presenting its material.

Another way this book differs from the others is that it focuses on the empirical tools of research before addressing the topic of forming research

xii INtroDUCtIoN

questions and hypotheses. I've often wondered when teaching my own courses out of other books, exactly how can a student choose an effective research question if they do not yet know *how* to model that question? More importantly, how can a student decide on a question if they do not know what statistical inference is? I have found that students who have learned or refreshed themselves with basic modeling and statistical inference early in class choose more sophisticated questions and hypotheses on which to perform their research, conduct better literature reviews regarding that research, and tend to write better papers overall. This, in my opinion, is simply because the rest of the semester is grounded *within* the context of empirical inference, which is the tool with which they will ultimately conduct their work. Books that focus on empirical work late in the curriculum force the student to design their research agenda before

they know how to actually apply that agenda, effectively putting the cart before the horse.

Since I have taught many undergraduate and graduate courses specifically related to performing research in economics, the organization of this book reflects my own views on how to obtain the best results possible for relatively novice writers, and therefore the best papers possible from your students. I am certainly not indicating that your students will be able to write grade A papers after reading this book, but I do believe that at the least there will be a statistically significant improvement in their performance relative to what their performance would have been had they not read it.

## **General Outline**

The book starts with the basics of econometric modeling. The first chapter is dedicated to linear modeling, programming in Stata and Excel, and inference from linear models. At this point, I leave any additional concepts such as heteroskedasticity, dependency, logarithmic transformation, etc., to the instructor to teach. Having said that, one knows that a research paper at this level does not necessarily need to address these issues to be a well-written, first-time research paper in economics. And even though I use Stata and mention it as the preferred econometric software package, I probably don't cover enough programming in Stata to

INtroductIoN xiii

satisfy the more rigorous instructors. However, I purposely leave the door open for these instructors to include a more comprehensive programming portion in their class curriculum. This is simply because the book is so short and basic. It really should only take half to two-thirds of the semester to instruct out of this book, leaving sufficient time to add material such as programming and other misspecification issues like dependency, heteroskedasticity, etc., and still finish the curriculum before the semester's end.

Starting at such a basic level of econometrics may be redundant to some in higher performing undergraduate and graduate programs; I truly don't expect instructors in Harvard's economics department to teach out of this portion of the book. This is because these students would likely have already taken a couple of courses on econometrics proper, rendering at least the first chapter obsolete. But, it is probably the case at most schools that the curriculum sequencing we desire to implement for majors (i.e., passing an econometrics course prior to a seminar course) does not always work out that way. Nevertheless, this book moves on in the second chapter to a way of modeling and drawing inference that may not be taught in an econometrics course the way I teach it.

I focus heavily on the separation between what I term "statistical" omitted variable bias and "theoretical" omitted variable bias. Within the context of this book, statistical omitted variable bias results from leaving out a squared right-hand-side variable when one is needed. We know that if a linear model is used when the relationship is

quadratic, biased inference will result. In this section, I also focus on calculating maximums and minimums, and how these apply within relevant sample spaces. I place emphasis on this area simply because the vast majority of theoretical relationships in economics are non linear in nature mostly due to some sort of diminishing returns. On the other hand, theoretical omitted variable bias exists when some variable,  $z$ , is left out of the conditioning set altogether. This type of bias results purely as a function of a theoretical relationship that influences the correlation between the variable of interest and the dependent variable. I have found that students understand these concepts better when separated in this fashion. From here I take the reader into basic fixed-effects modeling using panel data, and even touch on within transformations.

xiv INtroDUCtIoN

The remainder of the book structure focuses on performing literature reviews, how to read research papers, drafting, redrafting, finalizing the text, and presentation. Having published dozens of articles myself, I mostly rely on what works well for just me. Yet, I also realize that the research I perform is many levels of sophistication higher than the research these students will perform. To that end, I focus on “tricks of the trade” that will help the student bypass the rather technical aspects of writing a paper for publication in a top research journal, yet produce enough information to earn a good grade in a research seminar course.

In the end, I believe this book outperforms the others on the market simply because it is so simple. The instructor will be able to expand on this material without sacrificing precious class time, and the student will be able to use this book as a practical guide to writing a successful term paper. One would hope that performing good basic research in this field will expand the marketability of the student when they graduate, and this book should go a long way in attaining that objective. With that in mind, I hope you enjoy it!

## **CHAPTER 1**

# Basic Regression Analysis and Inference

This book's approach to regression analysis is truly basic. We won't go into issues such as the sum of squared errors, heteroskedasticity, statistical hypothesis testing (as opposed to economic hypothesis testing), t-statistics, f-statistics, etc. While all of these issues are very important for more sophisticated circles, it's simply not necessary to perform economic research at its most basic level. I will, however, use both Microsoft Excel as well as the regression package Stata. Even though we will be performing basic econometrics that can all be executed with Excel, the reader is encouraged to use Stata as this software is very easy to use with the student version of Stata costing well under \$100. Once learned, Stata can become one of the most useful tools for performing any kind of empirical research after the student graduates college.

## What Is a Regression Model?

We start with what a regression is meant to do—it is to generate an averaging line within a set of observations that are simultaneously determined by two variables. As with all lines, this line will have a slope and an intercept. Most of us remember grade school and the formula

$$y = +b \quad (1.1)$$

whereby,  $y$  is one variable,  $x$  another variable, and  $m$  and  $b$  are the line's slope and intercept, respectively; we use very similar notations in regression analysis. We use something called a regression model that takes the form

$$y = +$$

$$a_0 + a_1 x + e \quad (1.2)$$

where  $y$  is usually called the dependent or left-hand-side variable,  $x$  the independent or right-hand-side variable,  $a_0$  the line's intercept, and  $a_1$  is the line's slope; in regression jargon  $a_0$  and  $a_1$  are known as the estimators or coefficients. The new term in (1.2) which is not in (1.1) is  $e$ . This is what makes a regression different from a mere line. This is because a regression generates an estimate of the relationship between  $x$  and  $y$ , and the estimate itself is the averaging line. With any estimate there is a certain amount of error, and this is what  $e$  represents. So in essence we have an estimated line

$$y = + \hat{a}_0 + \hat{a}_1 x \quad (1.3)$$

with  $e$  being the difference between the actual relationship and the estimated relationship reflected in (1.3). The coefficients  $a_0$  and  $a_1$  are the estimates of the relationship's intercept and slope, respectively.

## Estimation and Inference

To make this clearer, assume we have two made-up variables, whereby  $x$  = yearly income of 10 home owners and  $y$  = each individual's respective cost of their homes—all in dollars. Table 1.1 lists these values and Figure 1.1 plots them on a graph. As expected, there is a positive correlation between annual income and the price of the houses these individuals purchased, that is, we expect that as incomes increase, people will purchase more expensive homes. Figure 1.2 shows the same plot as Figure 1.1, but with an averaging line through the dots on the graph.

**Table 1.1.**

**$y$  = House price (\$)    $x$  = Annual income (\$)**

427,900	147,000
78,500	54,600
66,950	47,325
247,900	92,500
110,500	67,450
85,000	51,230
399,000	111,250
410,900	119,500
156,500	72,400
105,900	60,750

To estimate the slope and intercept using *Excel*, one would

1. Open up an Excel spreadsheet and in the first cell of the first column and first row, enter a title for the income data (let's use *income* as the title for simplicity), and in the first cell of the second column, enter a title for the house price data (let's use *price* as the title).
2. In the subsequent cells underneath each title enter the data as shown in the table.
3. Click on the Data tab at the top and then click on Data Analysis; a pop-up box will appear. Scroll to Regression and click on it. Another pop-up box will appear.
4. With the cursor in the Input Y Range box, highlight the house price column including the title of the column. Move the cursor to the

600,000  
500,000  
400,000  
300,000  
200,000  
100,000

0 50,000 100,000 150,000  
**Annual income**

**Figure 1.1. Scatter plot of income on house prices.**

600,000  
500,000

400,000  
 300,000  
 200,000  
 100,000

0 50,000 100,000 150,000 Annual income

**Figure 1.2. Estimated relationship between income and house prices.**

Input X Range box, and highlight the income column including the title. Also, click the Labels box to the left of the Y and X input range boxes. We do this to tell Excel that the first cell contains letters, not numbers.

5. Click the OK button.

In *Stata*, one would complete the following steps in *Stata's* Do-file:

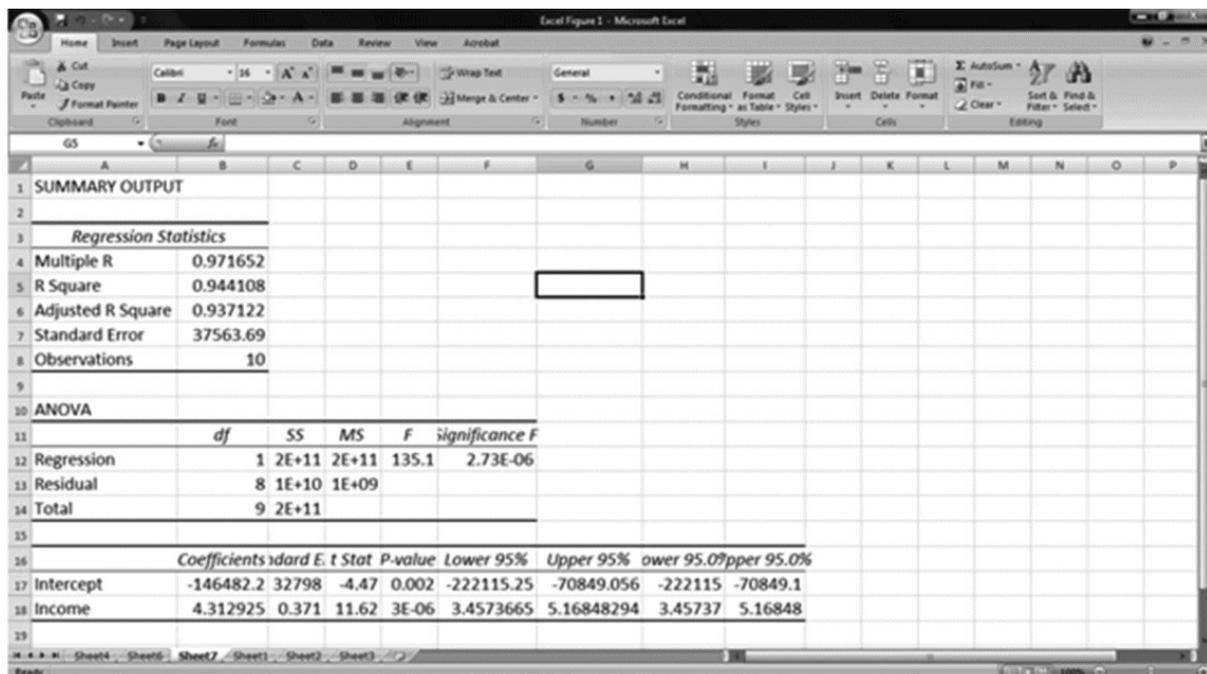
A. Perform steps 1 and 2 from the Excel instructions. Open *Stata* and click on the Data Editor (Edit) icon—it's the spreadsheet icon with a pencil in it. In your Excel spreadsheet, highlight your entire data set including titles and copy it. Then in your *Stata* Data Editor, click on Edit, click on Paste, and click on Treat First Row as Variable Names. Close the data editor.

B. Back in the *Stata* screen, click on Window, then on Do-file Editor and New Do-File; a blank pop-up window will appear.

C. Type in the following command: `reg price income`

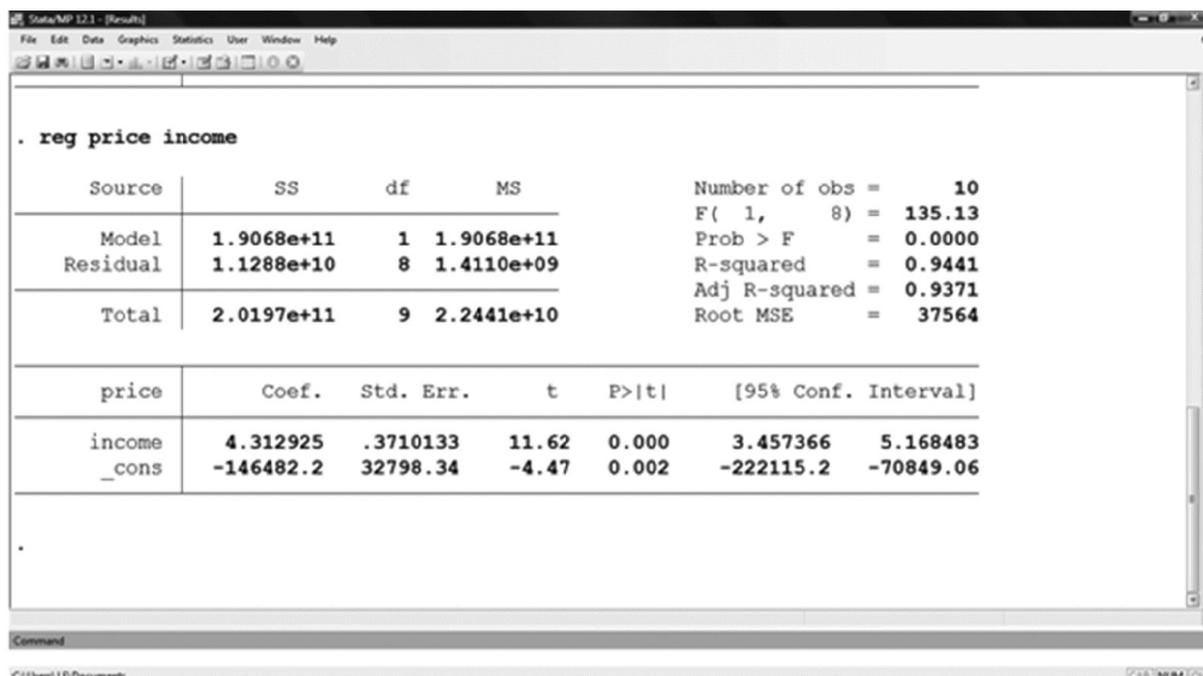
D. In the top bar of the Do-file Editor, click on the Execute (do) icon.

If you performed the Excel task, you will see



**Figure 1.3. Excel regression output for income.**

And with *Stata*, you will see



**Figure 1.4. Stata regression output for income.**

The reader can quickly see why Stata is the preferred package for regression analysis: it's simply easier to use and formats the output in a more readable way. Hence, from this point forward, we will only show Stata output, but will provide instructions to perform each operation in Excel as well. For now let us just focus on the numbers in the bottom two rows of figures 1.3 and 1.4; we'll get to what some of the other numbers mean later.

As we can see, the estimate of the intercept, `_cons` in Stata, is  $-\$146,482.20$ , and the estimate for the slope is  $4.312$ . Superficially one would conclude from this result that if incomes were zero, the average home price would be  $-\$146,482.20$ , and for every  $\$1000$  dollar increase in income level one would add  $\$4312$  to the price of a home. However, there aren't any home prices that are negative, at least to my knowledge. If one believes that house prices are at a minimum bounded below by zero, then one would have to resort to basic mathematics to complete this interpretation. Using the notation above, we can write this output in the form of a linear equation as  $Price = -\$146,482.20 + 4.312Income$ . We can set price equal to zero and solve for income to get  $Income = \$33,924$ ; in other words, on an average, individuals with incomes that are less than or equal to  $\$33,924$  per year will not own a home (or more precisely, own a home worth zero dollars).

Far more important is the interpretation of something known as a *marginal effect*. This concept is the focal point of nearly all economic research. Simply put, a marginal effect refers to a one-unit change in one variable causing some amount of change in another variable (this amount could be zero, or no change which we will address later in the book). Unfortunately, this is where we have to use a little basic calculus to determine the marginal effect.

Whenever someone hears the word *margin* used in the context of economics, one

should always think derivative (not financial derivative, but mathematical derivative). Readers of this book may be aware of basic economic concepts, such as marginal revenue, marginal cost, marginal product, etc. These are actually the derivatives of the revenue, total cost, and total product functions, respectively. Mathematically, one would represent a derivative of  $y$  with respect to  $x$  as  $\frac{dy}{dx}$ . (For those who are more mathematically inclined, I use partial derivative notation here simply to let instructors who want to go into interaction effects do so without having to change notation in their lecture(s).) To take a derivative of a function  $y = ax^r$ , we would perform the operation  $\frac{d}{dx} = rax^{r-1}$ . In our case, we have the function  $y = a_1x + a_0$ ; so performing the same operation on this equation would yield the derivative  $\frac{dy}{dx} = a_1$ . But since any variable to the power  $1-1$  equals 1, then we have  $\frac{dy}{dx} = a_1$ . Therefore, the slope of the regression line is the marginal effect!

In other words, the marginal effect of a \$1,000 increase (decrease) in annual income, on an average, will increase (decrease) the purchase price of someone's home by \$4,312. However, our inference is not yet complete.

Statistical inference requires two things, and sometimes three. At a minimum, inference requires an interpretation of the sign of the coefficient estimates (i.e., are they positive or negative), and analysis of the *statistical significance* of the estimates, especially the slope coefficients. The third component of inference, one which is not always addressed in research, is the magnitude of the coefficient estimate. Many times the estimate itself is less important than the sign of the estimate—an issue we will approach later in this section.

With regard to statistical significance, it may seem like the number 4.312 is a substantial distance from zero, but if there is a lot of variation in the data, we may not be *confident* that it is zero. To determine whether an estimate is significantly different from zero we will use something called *p-values*.

Interpreting a *p-value* takes some imagination. The data in Table 1.1 is a sample; it is not the entire population of income earners and respective home owners. And nearly all data one collects has this property. So assume the population consists of 100 home owners/income earners and we pull a sample of 10 observations from that population.

We run the regression we just ran, and get estimates of the coefficients equal to what we just got. Now, put those observations back into the population, and randomly draw another 10 observations. Run a new regression. The estimates you get from this regression may be similar but not the same because most likely the sample observations you drew this time were not identical to those drawn earlier. Then replace these observations back into the population. Draw a new sample and run another regression. The same thing will occur—you will get a new estimate of the slope and intercept that will likely be different from the first two sets of estimates for the same reason. If you do it 1,000 times, say, you will get a distribution of estimates of the slope and intercept. This distribution should resemble a bell-shaped curve with the center of the bell being concentrated around the value of the coefficient had one used the population instead of a bunch of samples (see Figure 1.5). Furthermore, as with any continuous variable, these values may span the entire real number line. The values of the estimated slope we obtained in the previous paragraphs could very well have been negative. Therefore, statistical significance must be calculated as a two-sided determination within a probability distribution.

To be 90% confident that the value you obtained is significantly different from zero results from determining whether 95% of the draws out of the 1,000 total draws, does not cross zero in value. We can only be 90% confident, and not 95%, because theoretically the estimate can take on any value on the real number line; if the distribution was on the negative side of zero then the upper tail of the distribution would come into consideration as well. View Figure 1.5 and imagine that if we were to draw a sample 1,000 times, would we get a positive (negative) value of our slope coefficient at least 950 times? If so, then I can be at least 90% confident that that estimate is in fact not zero. This would equate to a  $p$ -value of no more than 0.100. The  $p$ -value would equal 0.100 because if the probability of the coefficient taking on any value is 100%, and we have a 5% cutoff in each tail of the distribution, then  $100\% - 5\% - 5\% = 90\%$ ; 90% is the minimum amount of confidence we would have that the estimate is not zero. Since the  $p$ -value tells us the amount of the distribution that lies in the tails, and in this case no more than 10% of the estimates would, the  $p$ -value would be no more than 0.100.

As another example, assume that we again make 1,000 draws, but we only get a positive value of the slope coefficient 925 times, meaning that 75 estimates were negative. If the distribution was mostly on the negative side of zero, the upper tail would also hold 75 positive values. In either case, zero would lie *inside* the respective bounds of the 90% confidence interval in Figure 1.5. Since 75 is 7.5% of 1,000, then we could only be  $100\% - 7.5\% - 7.5\% = 85\%$  confident that the estimate is not zero. In this case the  $p$ -value would equal 0.150, or 15% written in percentage rather than decimals.

Hence, the lower the  $p$ -value, the more confident I can be that the coefficient estimate is not in fact zero. We call this a *statistically significant coefficient* or a *statistically significant relationship*. Our cutoff for the remainder of the book will be 90% confidence, or a  $p$ -value of 0.100 or less.

Value of coefficient

For positive values, if zero falls in this region, estimate is “significant”

For negative values, if zero falls in this region, estimate is “significant”

90%

**Figure 1.5. Coefficient distribution.**

From the values in Figure 1.4, we see that the slope coefficient estimate has a  $p$ -value of 0.000, and the intercept has a  $p$ -value of 0.002. This means that we can be 100% and 99.8% confident, respectively, that these values are not zero. We would consider these estimates to be statistically significant. If, for instance, the slope coefficient was not significant indicated by a  $p$ -value greater than 0.100, then we could not make the statement that incomes influence purchase price. In fact, we would then conclude that incomes “have no statistically significant effect on the purchase price of a home.”

In its totality, the full inference we would draw from this regression is that as incomes rise, so does the purchase price of the home that income earners buy. Furthermore, for every \$1,000 increase in incomes, purchase price rises by \$4,132. But one must ask whether this result would be the same if we controlled for other variables that influence someone’s purchase of a home, such as the number of children in a family? We can certainly assume that the more children in a family, the bigger house one needs and the more likely someone is to purchase a larger, and probably more expensive home. What if we held the number of children constant, would the marginal effect of income on house price be as large? We should add a variable to the regression that reflects the number of children for each individual in our sample and find out.

Table 1.2 is the same as Table 1.1, but with an additional column that lists the number of children for each individual sampled; for instance, a sampled person who makes \$92,500 per year with three kids purchased

**Table 1.2.**

$y$	= House	$x$	= Annual	$z$	= Number	price (\$)	income (\$)	of children
427,900		147,000		3				
78,500		54,600		0				
66,950		47,325		1				
247,900		92,500		3				
110,500		67,450		1				
85,000		51,230		0				
399,000		111,250		4				
410,900		119,500		5				

156,500 72,400 1  
 105,900 60,750 0

a house valued at \$247,900. We still have a linear regression model, but now it looks like

$$y = a_0 + a_1x + a_2z$$

whereby  $z$  is the number of children in the household.

Running a regression like we did before but now highlighting both the income and children column for the  $x$  input range in Excel, or for Stata, adding children to the command in line (c), we get the following output

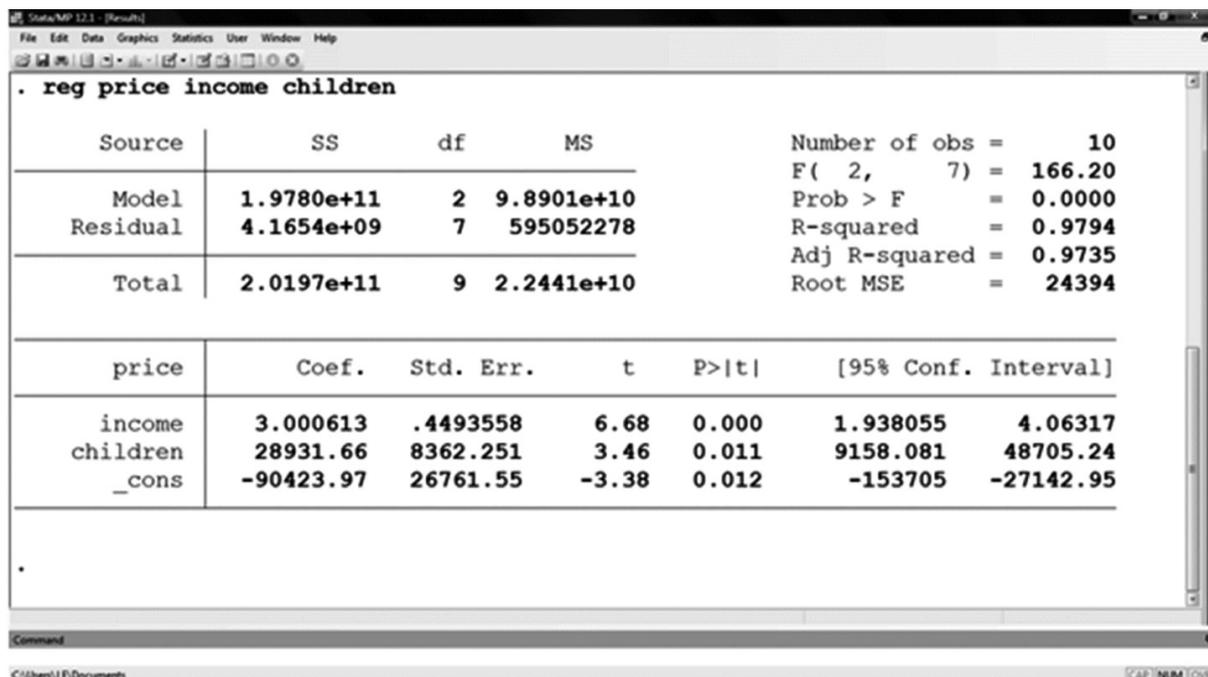


Figure 1.6. Regression output for income and number of children.

Drawing inference from these results we find that, as expected, holding the number of children constant substantially changes the marginal effect of income on the price someone will pay for a home. Now for every \$1,000 increase in income, house price increases by only \$3,000, not \$4,132—this result is also a statistically significant result like the previous one. We can draw inference from the estimate of the coefficient for the children variable as well. It is significant with a  $p$ -value of 0.011, and can be interpreted as: on an average, each additional child in the household increases the purchase price of a house by \$28,931, with holding income constant.

However, we cannot determine how expensive a home someone making \$100,000 per year would purchase without placing an actual value to the number of children. This is because the interpretation of the intercept is not the same as it was when there was only one variable on the righthand side. While the previous results could be graphed in two dimensions, as shown in Figure 1.2, and therefore we could simply look at the graph and guess what price this individual would pay for their home, in this case we

would have to graph this relationship in three dimensions. Here, the intercept is actually a common intercept for both relationships between income and price, and children and price. To interpret this outcome like we did before, we would have to say something like “an individual making \$100,000 per year in income, with 2 children, would purchase

a house equal to

$$-+3.000 *100,000+28,931* 2 =$$

90,423 \$267,439. The same individual with no children would purchase a home priced at

$$-+3.000 *100,000+28,931* 0 =$$

90,423 \$209,584. And imagine how

problematic this interpretation would become if there were more than two right-hand-side variables. These interpretation issues are precisely why many researchers stop inference at statistical significance and sign of the marginal effect and do not comment on the calculation of the estimated dependent variable.

At this point, I have said all I can about drawing inference on the estimated marginal effect. There is one other statistic that some consider important. Out of all the numbers in Tables 1.4 and 1.6, truly the only other bit of information that may be needed for basic research in economics would be the adjusted *R*-squared value. The *R*-squared is formally known as a sample correlation coefficient and tells the researcher how well their entire model fits the data. Specifically, it is a ratio of the variation in *y* that is explained by the entire set of right-hand-side variables, to the total variation in *y*. The model that generated the output for Figure 1.4 explains 93% of the variation in house prices, while the addition of the children variable adds roughly another 4% to that number.

We use the adjusted *R*-squared value instead of the nonadjusted value because we want to discount the *R*-squared by the number of right-hand-side variables we have in our model. Since no coefficient value is actually zero, even if it has a very large *p*-value, adding additional *x*'s to our model would increase its explanatory power of *y*. Therefore we need to adjust for this and the adjusted *R*-squared does exactly that.

I mentioned earlier that reflecting on the *R*-square value “may” be important information for a researcher. But to be honest, most good researchers ignore it. The reason is simple. Is a model with an *R*-squared value of 0.90 any better than one with an *R*-squared value of 0.30? Only in one instance—when forecasting a time series process. Model fit is critical when one wants to predict what inflation will be in the United States next quarter; but it is meaningless in the vast majority of economic research where the marginal effect is what is important. Especially for a cross-country growth researcher such as myself, low *R*-squared values are a common outcome; but, it doesn't mean that my model is a “bad” model. It simply means that there are many

factors outside of my variable set that explain the dependent variable—variables I probably don't have access to, or are simply not relevant to the research question I am asking.

### **Suggested Readings**

For an introduction to other aspects of econometrics, one should read

Naghshpour, S. (2012). *Regression for economics*. New York, NY: Business Expert Press.

Gujarati, D. N. (2003). *Basic econometrics*. New York, NY: McGraw-Hill Publishing.

For a more comprehensive introduction to Stata programming, one should read

Baum, C. F. (2006). *An introduction to modern econometrics using Stata*. College Station, TX: Stata Press.

Baum, C. F. (2009). *An introduction to Stata programming*. College Station, TX: Stata Press.

## **CHAPTER 2**