Bivariate & Multivariate Regressions: A Primer

By
Mark Zachary Taylor
Sam Nunn School of International Affairs
Georgia Institute of Technology
Original: Dec 2005
Current Version: November 2009

[Please cite as "Regression Analysis" in
21st Century Political Science: A Reference Handbook
edited by John Ishiyama and Marijke Breuning, (Sage Publications, 2010)]

I. Introduction

This paper provides a brief introduction to regression analysis. Regressions are a form of statistical analysis frequently used to test causal hypotheses in social science research. The simplest way of thinking about them is: given a scatterplot of data, regressions try to fit the "best" line to run through, and thereby describe, that data. Regressions are useful because that line can tell us a lot more about whether the data support a hypothesis than just the scatterplot alone. This probably sounds mysterious and unclear; the job of this paper is to demystify regressions.

Regressions are presented in research papers and articles that have a typical format. First the author poses a research question or causal hypothesis. She then reviews recent debate and research regarding this question. Next she suggests a statistical regression model and data with which to help answer the research question (i.e. test the causal hypothesis) and thereby advance the scientific debate. The big scientific payoff comes in discussing the results of the statistical analysis, which usually includes a few tables of regression results. Finally, conclusions are drawn, and perhaps implications are suggested for policy or future research. This is precisely the type of project you might have to write for a research methods course or senior thesis.

But how does one even begin to read regression results tables (such as Table 3 below on page 13), much less produce them? How does one judge whether a regression constitutes good evidence for a hypothesis? This paper will start you on the path towards answering these questions. In it, I attempt to do three things: 1) provide a basic explanation of what bivariate and multivariate regressions are and how they work; 2) show how regression results are reported in research articles, and explain how to read and interpret these results; 3) discuss briefly the conditions under which regression analysis can fail, and the techniques researchers use to address these failures.

These are precisely the skills you might learn in an introductory econometrics or intermediate statistics course. However, these courses are often taught by statisticians. And the difficulty with statisticians is that they often communicate in cryptic mathematical Greek symbols. Hence, when students walk into these types of courses, they are usually shown chalkboard after chalkboard full of mathematical derivations and proofs. A semester later, they walk out having no idea what it all meant, how it all fits together, or how they would actually produce or consume regression analyses.

So the idea behind this paper is to teach some fundamental concepts of regressions, such that you can walk into a regressions class and understand what's going on and what the statisticians mean by all of their Greek. Better yet, after reading this paper, you should be able to pick up a book on regressions and start teaching yourself.

II. The Good News About Regressions

Many of us start out a bit frightened of statistics, and with good reason. From our high school physics and chemistry courses, we know that math is very important for modeling, explaining, and predicting nature. We also tend to find in physical science courses that the more advanced or complex the math, the better it works. For example, in high school physics, we begin by using simple algebra; then in college physics, students learn to apply basic calculus; later, differential equations. Along the way, the simpler mathematics are discarded and dismissed as inferior. This is frightening since it means that the math gets ever more complicated.

The good news is that the opposite is true when applying statistics to study politics. First, in practice, simpler and more transparent statistical methods are often more convincing and more respected than are more complex or difficult statistical methods. The basic method of calculating regressions is called Ordinary Least Squares (OLS), which is what this paper will focus on. OLS is the "four-door sedan" of statistics. It is the technique that most researchers prefer to use. Certainly there are more complex and esoteric techniques, such as probit, logit, scobit, distributed lag models, panel regression, etc.[1] However, we only use these techniques in those special cases where OLS does not work so well. We generally prefer OLS where possible. In fact, a researcher who uses more complex techniques when OLS will do, or who does so for purely mathematical reasons, rather than pragmatic scientific reasons, often falls under suspicion.

The second piece of good news is that, when it comes to actually using statistics in research, understanding the mathematics is less important than understanding the concepts they represent. That is, a clear understanding of what regressions are and how to apply them can be more valuable than memorizing a bunch of mathematical proofs or knowing how to derive them. So what's the math good for? The math is very useful for answering questions we might have about a particular statistical method (e.g. is it appropriate for this type of data or that kind of situation?); the math allows us to discuss what is going on in a very precise way. It's a little like learning auto mechanics versus learning to drive. Being a good mechanic can be a life-saver if your car breaks down. And certainly, if you want to be a professional racer, then you'd better know your mechanics. However, you do not necessarily need to be able to take apart an engine in order to drive a car. So, if you have a firm grasp of the fundamental concepts of regressions, then you do not need to be a whiz at the mathematical mechanics. On the other hand, if you do not understand the fundamental concepts, then it does not matter how great you are at the math, your statistics will be useless. You might scare some critics off with your mathematical dexterity, but researchers who know their fundamentals will be able to see right through you.

III. Drawing a Regression Line: The Basics of Bivariate Regression

Given knowledge of introductory statistics (i.e. descriptive stats, probability, statistical inference), a student's next step is typically to take intermediate statistics, which for political scientists is always regressions. Why? Well, in social science, one of our main goals is to test theories of causality. We hypothesize that X causes Y; then, to test this hypothesis, we gather data and use regression analysis to see whether that data shows any evidence of a causal relationship between X and Y.

Couldn't we use introductory statistics to ask whether X and Y correlate or covary? Sure we could. But it wouldn't tell us a whole lot. Let's see why with an example of a bivariate regression, which is a regression that has only two variables: one independent and one dependent.

For example, say the city newspaper reports that a strange flu has broken out in town, and it prints a chart of the number of sick people by ZIP code. You talk to a few flu-sufferers and ask what they did during the days leading up to the flu. You find that they have one thing in common, they each went to meet the city's mayor. So you ask yourself: why should that cause the flu? Then you realize that the mayor is one of those old-style, glad-handing politicians who likes to handshake and hug everyone she meets. So you hypothesize that the flu is caused by a virus which the mayor has and he's passing it on by this contact. That's gotta be it! But if you called the mayor's office or state health officials with this claim, they would think you are crazy. To better convince them, you need to provide some evidence to support your hypothesis. How can you do this?

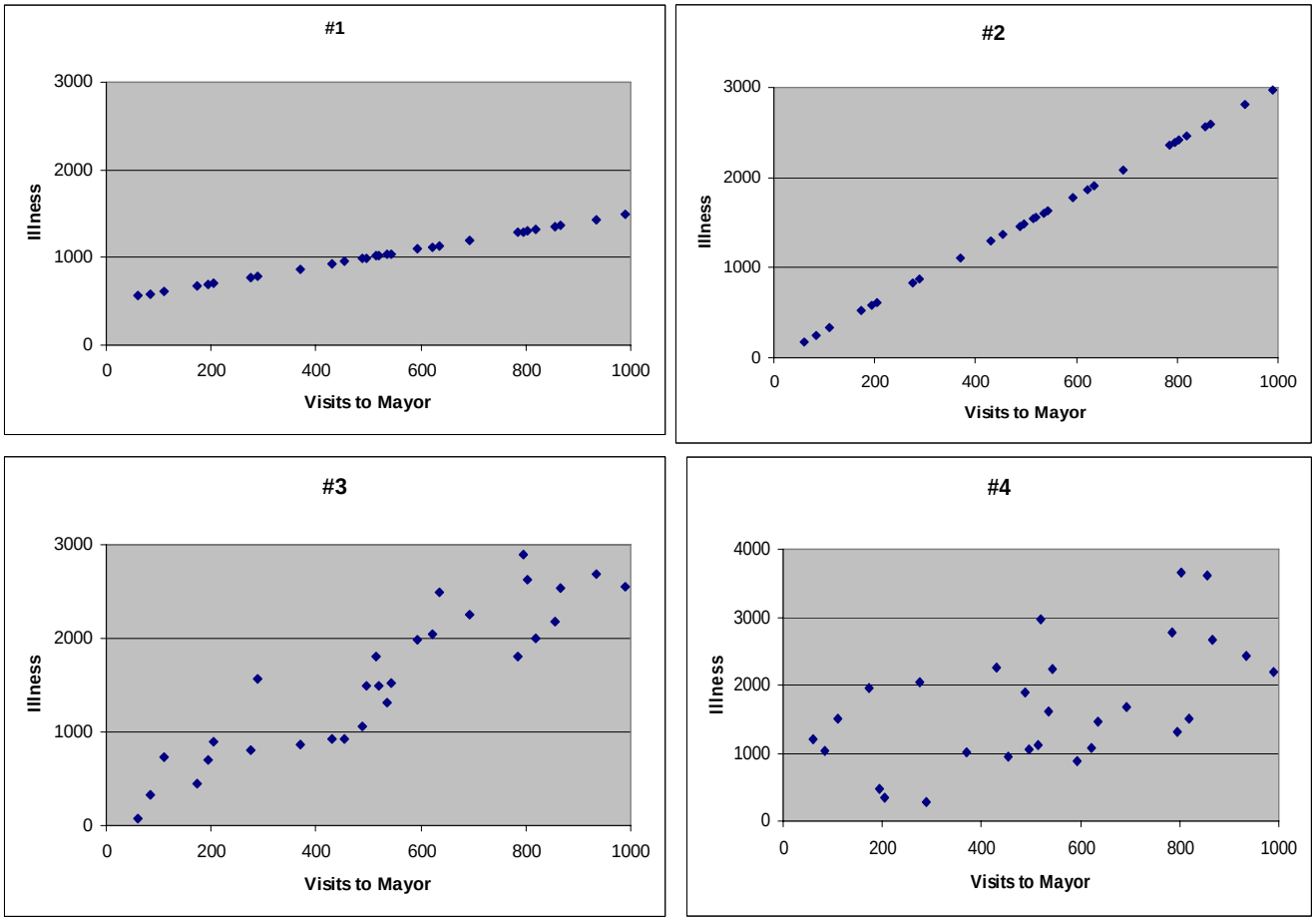
You certainly cannot contact everyone in the city, asking if they visited the mayor recently and now have the flu. However, you could consult the mayor's official visitors sign-in book which records each visitor's name and address. So for each ZIP code, you could plot the number of visitors to the mayor on one axis and, using the newspaper's chart, the incidence of flu on the other. It's simple, just draw a scatterplot!

Figure 1 shows a few possible scatterplots that might result from this data. Each scatterplot represents just one of an infinite set of possible results. If your data produced any of the first three scatterplots, then you could be pretty sure that there is a relationship between visits to the mayor's office and incidence of the flu. But in the fourth scatterplot, you are not so sure. Therefore the first thing you would like is a technique by which you could be more confident of whether you are observing a relationship or not in scatterplot #4. Descriptive

[1] Probit, logit, and scobit are used when the dependent variable is not continuous. Distributed lag models are used to analyze variables that change over time and where the current value of the dependent variable is partly explained by its previous ("lagged") values. Panel regressions are used to analyze variables that change across both time and space (e.g. country, state, city).

statistics, such as covariation and correlation, will not help you much with this.

Figure 1: Illness vs. Visits to the Mayor



Also, look at the first three scatterplots. Each graph clearly shows some sort of a relationship between visits to the mayor and incidence of flu, but they are distinctly different relationships. It would be nice if you could say something about *how* they are different. Again, descriptive statistics are not much help here.

Let's start by figuring out how the relationships are different in scatterplots #1, #2, and #3. You do not have to go much further than high school algebra for help. In order to see how these relationships are different, you can simply draw a line through the data

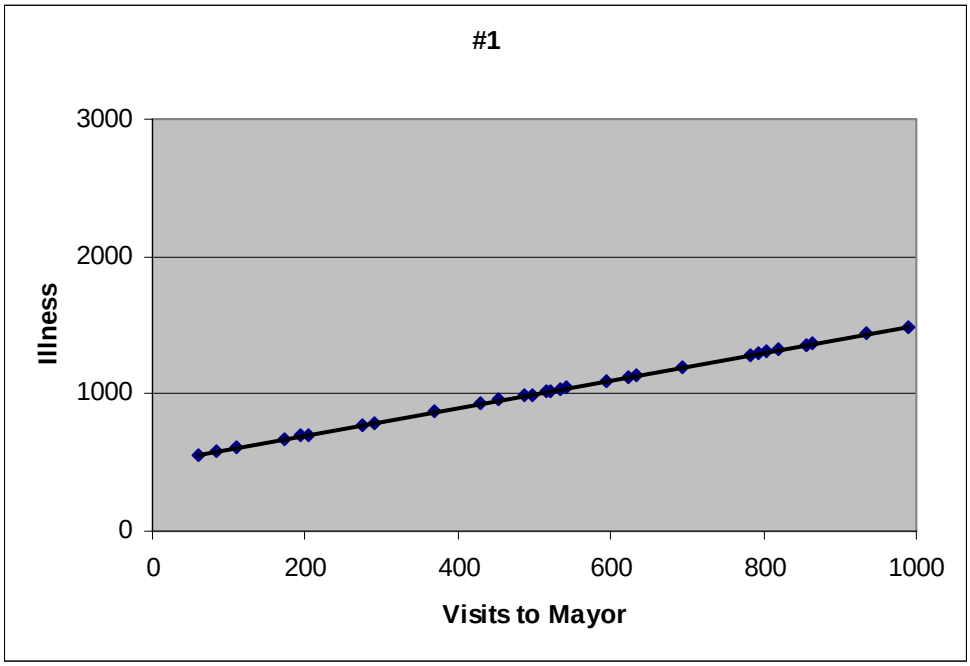
How does drawing a line help? Remember from elementary algebra that, given a bunch of Y's and X's defined as points on a line, the equation for that line is:

Y = (slope * X) + intercept

First, the slope of the line tells you how much of an increase in Y is related to an increase in one unit of X. Second, the intercept of the line with the Y axis tells you how much Y there will be when there is no X at all (when X = 0).

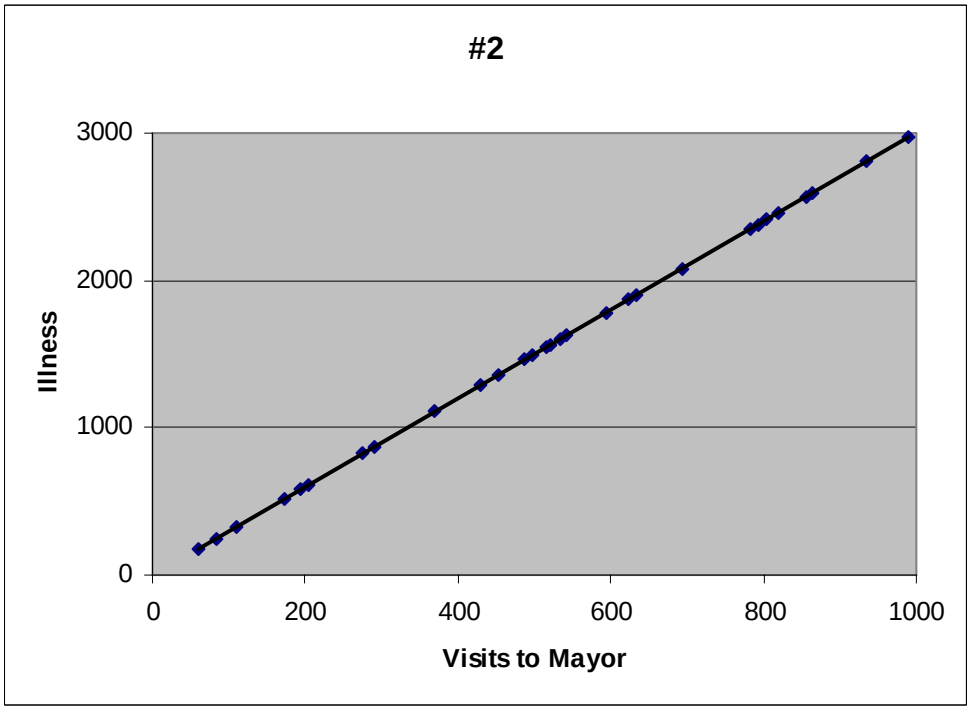
So let's try to draw some lines through your mayor visits vs. illness data. The first two are easy since the datapoints line up perfectly. In scatterplot #1, you would use simple algebra on the raw data (not shown) to figure out that the slope is 1, the intercept is 500, therefore:

Scatterplot #1: (sick people = 1 * [mayor visits] + 500)



In scatterplot #2, we again use the raw data to calculate that the slope is 3, the intercept is 0, therefore:

Scatterplot #2: (sick people = 2 * [mayor visits] + 0).



Now that we know these simple equations, we have two drastically different interpretations of the data. In scatterplot #1, the slope is 1. This means that, for every one visit to the mayor, one person gets sick. Also, the intercept is 500; this implies that, if no one visited the mayor, then 500 people would be sick. So the line in scatterplot #1 does *not* look like a flu virus! That is, if someone visits the mayor and catch the flu, then he should bring it home to infect his friends, family, and then his coworkers. So every one visit to the mayor should probably result in multiple people getting sick. In other words, the slope should be higher, which is exactly what

we see in scatterplot #2.

Furthermore, if the mayor is the source of the flu then the following should apply: if no one visits the mayor, then few people should get sick. That is, the intercept term should be close to zero. Again, this looks more like what we see in scatterplot #2. But in scatterplot #1, the line predicts 500 sick people even when no one visits the mayor.

Therefore the equation for the line drawn in scatterplot #2 looks like evidence of a virus: if no one visits the mayor, no one gets sick; and for every person who visits the mayor, several people get sick. Meanwhile, the equation for the line drawn in scatterplot #1 *does* show that visits to the mayor's office correlates with illness, but this relationship *does not* look like a virus. So if your research produced scatterplot #1, then you might have to reject the virus hypothesis and formulate another one. For example, maybe it's the bad coffee they serve, or mold spores from the old walls in city hall, or some sort of noxious gas at that particular subway stop. In this case, you would then want to gather new data (e.g. on the incidence of coffee drinking, mold allergies, etc.) to better test these new hypotheses.

This is exactly how political scientists use regressions. We first ask "what causes Y to vary?"[2] Then we formulate a hypothesis in which we theorize an X that causes Y to vary.[3] Then we gather data on X and Y, and use regression analysis to draw a line through the data. Finally, we ask whether the slope, the intercept, and perhaps the shape of the line supports our hypothesis about what's going on between X and Y.

Now let's look at scatterplot #3 again. The data points do not line up exactly. That is, for any line we might draw, most points will not lie on it, they will "wander" away from the line. So, how do we draw a line through this cloud of data points? We could draw a line with a high slope and zero intercept (like a virus), but we could just as easily draw a line with a low slope and positive intercept (like bad coffee). Come to think of it, we could draw lots of different lines through this data. But what's the best, most honest and objective line that we can fit to this data?

The answer is to minimize the "errors". Here "error" does *not* mean a mistake or that something has gone wrong with the data. Rather, "errors" are the vertical spread of the data around the estimate (i.e. the regression line), hence they are also referred to as the "residuals". One way to remember this is that the term "error" is derived from the Latin word *errare* which means "to wander". So when you hear about errors, think of it as the amount by which the datapoints "wander" away from the regression line (see Figure 2 below).

So what's the best way to minimize the errors? It turns out that the best way to fit a line to data is to minimize the squares of the residuals. This takes care of a mathematical "positive-negative" problem of having errors both above and below the regression line. It also weights the data such that points further away from the line bring a heavier penalty than those closer to it. This technique (Figure 3 below) essentially draws an ordinary square connecting each data point to the line, and uses mathematics to ask "what is the collection of smallest (or 'least') squares we can find": Ordinary Least Squares (OLS).

At this point in a typical regressions class, the next step would be to introduce the Gauss-Markov Theorem, which is the foundation of all regression analysis. The Gauss-Markov Theorem proves mathematically that OLS always fits the best lines to data, at least under certain conditions. These conditions are called the Gauss-Markov assumptions. Usually a significant amount of any research paper using regressions involves explaining whether or not these assumptions hold, and what "fixes" the researcher has applied if they do not. Understanding these assumptions, and what to do when they do not apply, is also what students do in the second half of a typical regressions class, which we will discuss later in this paper.

---

[2] It is often a better strategy to begin research with a question about the *dependent* variable and what affects it. Why? Because any given *independent* variable can have an infinite number of effects on the world. Hence, research that begins with questions about the effects of an *independent* variable can often lead even experienced scholars either on wild goose chases (unfocused research) or to purely descriptive and biased answers (no scientific method, just reporting). So be careful when designing your research question to stay focused on the dependent variable.

[3] Good researchers also include an explanation of the causal mechanism (precisely how X causes Y to vary) in their theories.

Figure 2: Regression Line and Errors for Scatterplot #3

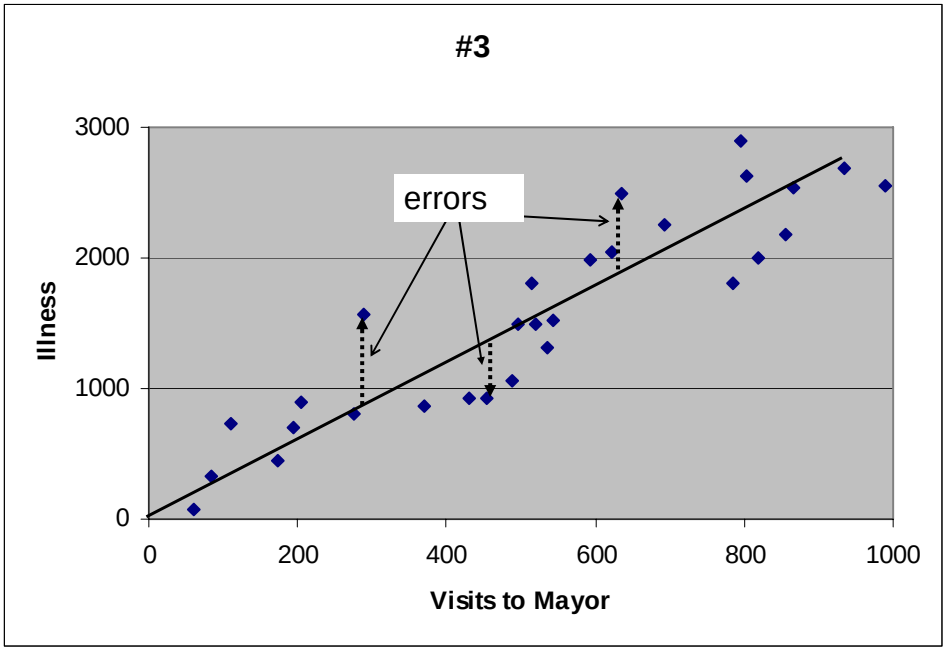
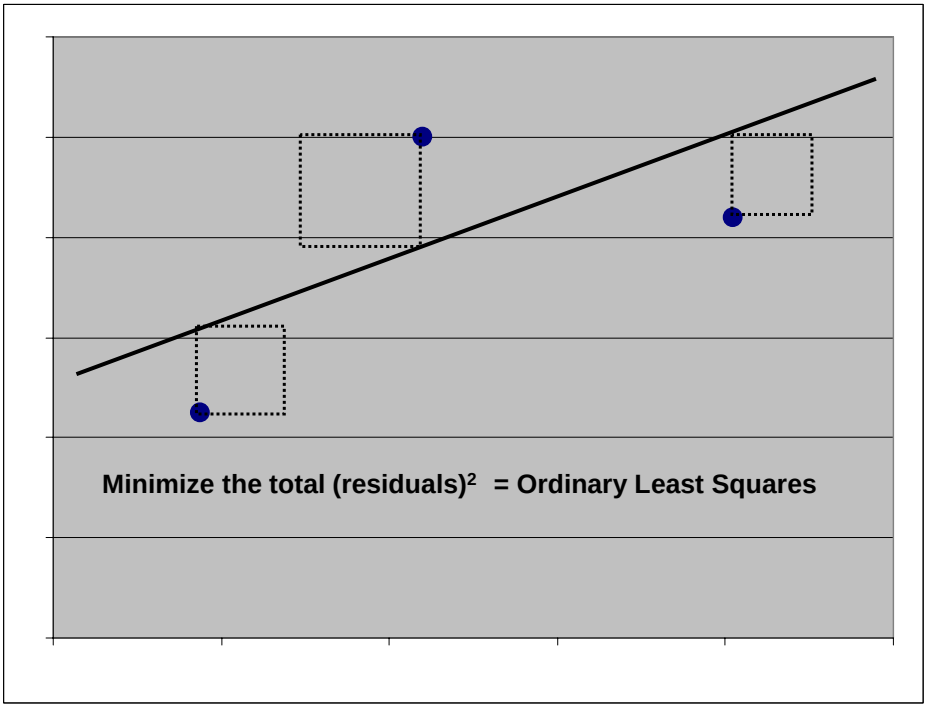


Figure 3: Ordinary Least Squares on Three Datapoints



Also in a typical regressions class, students might practice calculating some of the math by hand. But in practice, statistical software packages such as STATA, SPSS, SAS, R, Eviews, and dozens of others do the mathematical work for you. Data can be directly entered into these programs, or imported from a typical spreadsheet program.[4] For example, Table 1 shows the spreadsheet of data used to create scatterplot #3.

Table 1

[4] Many spreadsheet programs, such as Excel and CALC, can perform basic regressions as well.

Zip Code	Visits	Illness
30301	515	1813
30302	935	2681
30303	487	1062
30304	594	1986
30305	694	2256
30306	275	810
30307	865	2531
30308	454	930
30309	173	443
30310	520	1487
30311	60	79
30312	542	1516
30313	820	2003
30314	784	1804
30315	430	931
30316	804	2628
30317	193	708
30318	290	1562
30319	623	2050
30320	204	891
30321	497	1489
30322	989	2558
30323	634	2497
30324	370	859
30325	110	728
30326	83	323
30327	794	2895
30328	535	1316
30329	856	2173

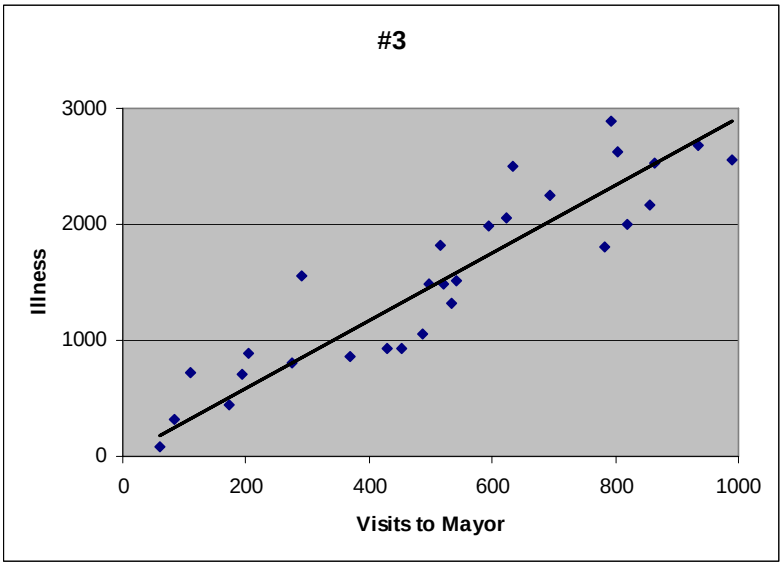
After creating this table, I imported this data into my statistical software package (STATA). To perform the regression, I then typed a simple command (in this case: "regress illness visits")[5], and received the computer readout below (Table 2) which is typical of that provided by many statistical software packages. This readout may look ugly but it's just computer-speak for Figure 4 (below). Let's now discuss how to make sense out of it.

Table 2: Sample Readout of Regression in Scatterplot #3

```
. regress illness visits
      Source |       SS       df       MS              Number of obs =      29
-------------+------------------------------           F(  1,    27) =  138.41
       Model |  14726524.1     1  14726524.1           Prob > F      =  0.0000
    Residual |  2872694.91    27  106396.108           R-squared     =  0.8368
-------------+------------------------------           Adj R-squared =  0.8307
       Total |    17599219    28  628543.534           Root MSE      =  326.18
------------------------------------------------------------------------------
     illness |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      visits |   2.682265   .2279892    11.76   0.000     2.214469     3.15006
       _cons |   152.6323   133.4816     1.14   0.263    -121.2493    426.5139
------------------------------------------------------------------------------
```

[5] Commands will vary across different software packages; my point here is merely to illustrate how simple statistical software packages can be..

Figure 4: Scatterplot #3 With Regression Line (y = 2.68x + 152)



First, where in the computer read-out (Table 2) is the line's equation? You can find it in the lower left quadrant of the readout. The "Y" or dependent variable (ILLNESS) is listed at the top of the column, with the independent variable (VISITS) and intercept term ("_cons" for constant) below it. The values of the slope and intercept term are listed in the column labeled "Coef." (for coefficient). Therefore, the readout tells us that the line which best fits the data is described by ILLNESS = 2.68 (VISITS) + 152. We get two pieces of information from this equation:

1) For every 1 additional visitor to the mayor, an average of 2.68 people get sick
2) When nobody visits the mayor, an average of 152 people get sick.

Note that we do not interpret these numbers as "1 visit causes 2.68 cases of illness". Why? Remember that regressions can show correlation but *not* causality. Social scientists use statistics to argue for causality by asking whether the correlations they show match those predicted by their hypotheses. Thus statistics rarely prove a hypothesis, but they can provide evidence for or against it. Hence, the researcher's most important task is to identify the regression equation and data that will best test his hypothesis, and thereby best convince an audience of fellow scientists that it's true.

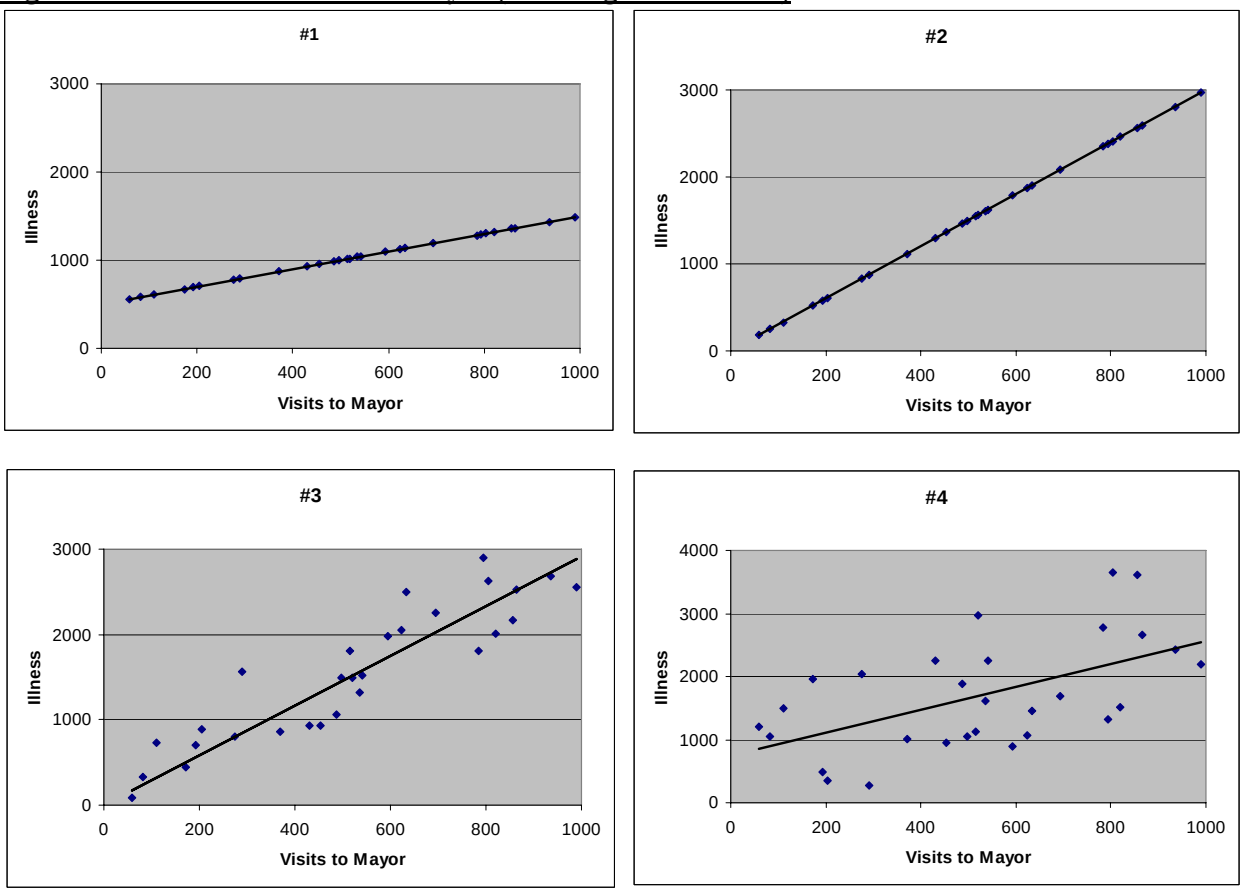
But what about all of the other information presented in the computer readout?

IV. The Coefficient of Determination (R^2)

R^2 or "R-squared" is known as the "coefficient of determination" or the "goodness of fit test". Its value can range from 0 to 1, it tells you how linear the data is. What does this mean exactly? Let's look again at our four scatterplots above, but this time with the OLS regression lines drawn through them (Figure 5 below). It may not look like it, but the relationships estimated in scatterplots #2, #3, and #4 are all based on the same underlying equation: y = 3x; hence the true regression line's slope is 3, the true intercept is 0. But there is something clearly different going on in each of them. Let's investigate.

In scatterplot #2 the line goes through each of the data points. That is, the regression line completely fits each and every data point; it therefore explains all the data. This means that all the variation in Y is explained by X; all the variation in ILLNESS is explained by VISITS to the mayor.

Figure 5: Illness vs. Visits to the Mayor (with Regression Lines)



But in scatterplots #3 and #4, the data points "wander" or "err" on either sign of the line. That is, the regression line explains some of the data (the variation in X explains some of the variation in Y), but there is some left over, unexplained, or "residual" variation. More precisely, for any datapoint not on the regression line, part of its height is explained by the line, part is not. The percentage of $(\text{height})^2$ that is explained by the regression model tells us the percentage of Y explained by X. That percentage is known as R^2 and is reported as a number from 0.00 to 1.00. In scatterplots #1 and #2, the $R^2 = 1.00$ because X explains all of the variation in Y; there is no unexplained portion of the data. In scatterplot #3, the $R^2 = 0.84$ which means that 84% of the variation in ILLNESS is explained by VISITS TO THE MAYOR. In scatterplot #4, the R2 = 0.31; which means that, for this data, 31% of the variation in ILLNESS is explained by VISITS TO THE MAYOR.

Put another way, in the same manner that OLS draws squares to fit the best line through the data, we can perform a similar operation for each data point. We can square the height that is explained by the regression line, and compare it to the square of the residual height. We can then add up all those squares and use them to get a sense of how much variation in Y the regression model explains. The following equation better illustrates this concept:[6]

Total Sum of Squares = Unexplained Sum of Squares + Explained Sum of Squares.
(a.k.a. error, residual) (a.k.a. model, regression)

What's behind the unexplained part of the data? It might simply be randomness. For example, people in one ZIP code may have been accidentally rubbing virus into their eyes more because of a random spike in air pollution this month, or the weather in another ZIP code was randomly warmer, making flu resistance slightly

[6] Warning: different books use conflicting acronyms to refer to these components. For example, ESS might be used to indicate "Error Sum of Squares" in one textbook, but "Explained Sum of Squares" in another. Likewise RSS might mean "Residual Sum of Squares" in one textbook, but ""Regression Sum of Squares" in another. Be aware of this when consulting different books on this topic.

higher there.

However, there could also be a systematic cause for some of the unexplained data. Let's say that Norwegians are genetically more susceptible to this mystery flu, and the neighborhoods represented by datapoints above the regression line have more Norwegians than those below. The ZIP codes with more Norwegians would then have higher incidence of flu than we could explain using just VISITS TO THE MAYOR. And if we included another X in the regression, in which X = "percentage of the population that's Norwegian", we would get a higher R^2 because that part of the variation in Y would be explained. We will talk a little more about adding more X's and what that means later. But first let's finish up with R^2.

Now look back at Table 2, you will find that we are beginning to understand more of the computer readout. In the upper left quadrant of the readout, you will see a small sub-table. In it, the SS ("sum of squares") is split into the three parts that we just discussed: 1) Model, 2) Residual, and 3) Total. Note that if you divide the Model by the Total you get R^2.

A small warning about R^2 is appropriate here. Decades ago, eager young researchers used to jump on this R^2 measure and conclude something like: "Aha! Regressions #1 & #2 are *better* than #3 because they have higher R^2; likewise regression #3 is *better* than #4 because it has a higher R^2. It explains more of the variation in the dependent variable." But this is not entirely accurate. Remember that the true equation that generated the data ($y = 3x$) is the same for scatterplots #2, #3, and #4. The only difference is that the latter scatterplots have higher levels of randomness added to the data. So my exact equation for these scatterplots is: $y = 3x + \varepsilon$, where "ε" is some small random number. But the fundamental relationship between X and Y is the same across the scatterplots, therefore one regression line is not somehow "better" than the other. Yes, in regressions #3 and #4 there is a large amount of unexplained variation that is not present in #2. But that does not necessarily make the former "bad" regression lines. After all, regressions #3 and #4 still show a relationship between X and Y that indicates that a virus is present: high slope, low intercept.

So what does R^2 really tell you? It does *not* tell you whether you have a relationship between X and Y, or whether one regression estimate is somehow better than another. What it tells you is how closely grouped around the regression line the data is. It tells you how *linear* the relationship is. For example, if we had a "U" shaped scatterplot, then OLS would produce a low R^2 since it would not be able to fit a straight line to the "U" of data. But that low R^2 does not mean that there is no relationship at all, just that there is no linear relationship. So, although you generally cannot use R^2 to judge whether one regression is better than another, you might use R^2 to judge whether the data is linear, whether there are some missing independent variables causing a lot of systematic error ("wandering"), or whether there is a large component of randomness affecting your dependent variable.

V. Standard Errors of Coefficients

The most important information to come out of any regression is often the slope of the regression line (a.k.a. the coefficient) and perhaps its intercept. These are the two pieces of analysis that really tell us something useful about the relationship between X and Y. For example, remember our four scatterplots above. In them, the slopes, and to a lesser extent the intercepts, provided evidence as to whether or not we were dealing with a contagious virus or not.

But the numbers that OLS produces for the slope and intercept are just statistical estimates. Again, recall that the city has a large population, split up into dozens and dozens of ZIP codes. In our example, we merely took a sample of them, and we used this small sample to estimate the unknown parameters of the entire population. But how confident can we be in our estimates? In our example, how confident can we be that the true relationship, the relationship for everyone in my city, is actually: ILLNESS = 3 (VISITS) + 0?

In other words, the slope we computed is just an estimate of the real value. But because of sampling variability, there is always a probability that this estimate may be a little high or a little low. But how high or low? One important indicator of how far off the estimate might be is the "standard error" of the slope coefficient. The standard error is simply the standard deviation of how much the data "wanders" around the regression line. It can therefore give us an indication of how much that point estimate is likely to vary from the true value.

You can use the standard error of the slope coefficient in the same way that you used standard deviations to judge sample averages in introductory statistics. In introductory statistics you should have learned that if a sample is selected at random, then as you increase the sample's size, the mean and variance of that

sample will look more and more like the mean and variance of the population from which that sample was drawn. Furthermore, we know that for data that are normally distributed, 68% of the data will lie within one standard deviation of the mean, 95% of the data will lie within two standard deviations, etc. And thanks to the Central Limit Theorem, we know that many important statistics, like means and variances, are normally distributed.

The slope and intercept terms that we estimate in regression analysis are no different. We can measure how much our data wanders, and use this information to get a sense of how accurate our estimates of the slope and intercept are. Due to the Central Limit Theorem, we can say that, yes the slope and intercept we estimate from our sample may be a little higher or lower than that of the population from which our sample came. But 95% of the time, the estimates based on our sample data will wander within plus or minus two standard deviations of the population's values. Thus we can be 95% confident that the population's slope and intercept lie within two standard errors (technically 1.96) of the coefficients we estimated from the sample.

For example, if our slope estimate is 3 and the standard error ("the standard deviation of the wandering of the data") is 0.5, then we can be 95% confident that the true slope is 3, plus or minus two standard errors. That is, the sample came from a population with a regression slope between 2 and 4. This means that we are 95% confident that there is a positive (i.e. non-zero) relationship between VISITS and ILLNESS. We therefore say that the coefficient on VISITS TO MAYOR is "statistically significant", given a 95% confidence level. In this particular case we can go further, and say that the slope is greater than 1, that this relationship is greater than 1-to-1 (between 2-to-1 and 4-to-1). Hence if a virus hypothesis predicts a relationship greater than 1-to-1, then our data supports this hypothesis.

However, if our slope estimate is 3 and its standard error is 2, then we can be 95% confident that the population's slope is between -1 and 5. Note that this range includes zero! In other words, the data wanders so much, and therefore our estimate of the slope wanders so much, that it probably wanders over zero. And if the slope is zero, then there is no relationship; we cannot confidently reject the possibility that there is *no* relationship between VISITS and ILLNESS. Hence, one of the most important implications of the standard errors is whether we can be 95% confident that the slope coefficient is *not* "0" (no relationship between X and Y). A quick rule of thumb here is that if the coefficient is greater than twice the standard error, then it is significant at the 95% confidence level.

There are four ways of reporting this information, all of which can be found in the lower table in the computer readout (Table 2). The first way is simply to report the standard errors, and let the reader do her own multiplication or division by two (technically 1.96). Second is to report the results of the division. That is, report the ratio between the coefficient and the standard errors. Remember that we want two or more standard deviations away from zero in order to be confident in rejecting it. So we want a ratio greater than two (technically 1.96). This is known as a t-score, t-ratio, or t-statistic. Third, you could report the p-value, which is the probability of the coefficient being zero. Finally, you could report the confidence interval itself.

Since these measures are just different ways of reporting the same data, there is no pressing reason for choosing one over the other when writing up your regression results. Authors usually select the measure preferred by their publisher, or most commonly used in their sub-field, or that they have inherited from their professors.[7] When reading regression tables, this means that you should check the small print, usually a footnote to the results table itself, to see which measure is being reported. And you should similarly include this information when writing your own regression results.

Furthermore, when reporting regression results in a table, authors usually highlight "statistically significant" findings with asterixes. That is, coefficients which the author is 95% confident are not zero might receive a single asterisk next to the standard error; coefficients with 99% confidence get two asterisks, etc. For example, a coefficient of 3 with a standard error of 1.5 would have a t-statistic of 2, or a p-value of 0.05; and regardless of which of these were reported, it would receive an asterisk next to it.

Of course, you cannot simply analyze regression results mechanically. It is very important that you develop your own sense of judgment about which findings are important. Take for example the 95% confidence level. It is an entirely arbitrary number, dictated more by tradition than anything else. There is no mathematical or scientific justification for choosing 95% rather than 90% or 99%. Indeed, one could argue that in some cases

[7] The confidence interval, however, is rarely reported in published research articles since the other expressions of error give more precise information.

80% might do just fine. Why? The less data included in a regression, the more difficult it is to distinguish statistically whether a coefficient is significant. Small deviations of individual datapoints could easily change the slope estimate of the line between them. Therefore if you performed a regression on only a handful of datapoints, a lower confidence level might be justifiable. On the other hand, if you had a few thousand datapoints or more, then you might require a higher confidence level for significance, say 99.9% or higher.

Finally, another type of judgment you will need to develop is about how to balance significance with the coefficient itself. For example, say you run a regression which produces a tiny coefficient that is highly significant. Sure you will have a finding, but it will be a tiny finding. So it might be statistically significant, but not substantively significant: your X will have a tiny effect on Y. Now consider the opposite case in which you find a coefficient that is large but insignificant. Here you have an absolutely huge (and therefore substantive) effect of X on Y, but it is not within the arbitrary confidence level. So it may still be worth reporting as an interesting result, worthy of further study.

VI. Multivariate Regressions

Bivariate regressions are useful, but usually when we want to explain something, we have more than one independent variable that we want to control for. Let's go back to our flu example. What if we finally realize that the number of Norwegians in a ZIP code affects how many people there get the flu, or maybe we want to control for whether the district gave out flu shots, or access to health care, etc. In physics and engineering, when you start adding more variables, things start getting really complicated, as do the mathematics to explain them. Good news: not in statistics! Regressions work almost exactly the same with two variables as with three, four, or one-hundred. You do have to think a little outside the box though.

So far, we have found that bivariate regressions are an improvement over basic statistics. We can use them to calculate the slope and intercept of a regression line. The regression line tells us much more about the relationship between two variables than if we had just compared their means or discovered that they covary. But the real payoff of regression analysis comes when we move from the bivariate case of X causing Y, to the multivariate case of two or more different X's causing Y.

Why? Put simply, the coefficients produced by multivariate OLS tell us the amount that Y changes for each unit increase in each X *holding each of other independent variables constant*. This ability to "hold constant" is hugely important! In controlled laboratory experiments, scientists hold everything constant except for one causal variable. They then vary that one causal variable, and see what effect it has on the phenomenon they are studying (the dependent variable). It is their ability to isolate all causal factors except one that gives controlled experiments their great explanatory power!

Multivariate OLS allows us to hold variables constant mathematically. Say we want to study the effect of two independent variables (X_1 and X_2) on a dependent variable (Y). When calculating the effect of a change in X_1 on the average Y, OLS mathematically partials out the effect of X_2 on Y. It also mathematically strips out the parts of X_2 that might affect X_1. This has the effect of isolating X_1 and its effect on Y. Hence we can interpret the coefficient on X_1, as being the average change in Y for a unit change in X_1 holding all other variables constant (or controlling for all other variables). Of course, OLS simultaneously does this for each of the other independent variables included in the regression. So the coefficient for each independent variable can be interpreted as the effect of that variable, holding all of the others constant. This is fantastic for political scientists since it is usually not practical or possible for us to conduct laboratory experiments. It means that in situations where we cannot exert experimental control to produce data and thereby test hypotheses, we can instead use OLS to exert statistical control over data we collect and thereby test hypotheses.[8]

In the multivariate case, we still have just one Y (just one dependent variable, just one effect we are trying to explain), but now we can estimate the effects of multiple different X's (multiple different causal variables). The concepts all work exactly the same way as for the bivariate case. The basic difference is the number of dimensions. In the bivariate case, we have a two-dimensional plane of data (x on the x-axis, y on the y-axis), and OLS fits a one-dimensional line through it. When we increase to the three variable case (two independent variables and one dependent variable), then we have a three-dimensional cube of data, and OLS fits a two-dimensional plane through it. But we still interpret the coefficients and standard errors just the same as in

[8] For those of you who need more convincing, and want to see exactly how OLS exerts mathematical control over data, I refer you to the books recommended at the end of this paper.

the bivariate case.

For a quick example, let's go back to our mysterious flu case and say that we suspect Norwegians are much more susceptible to this flu than everyone else. So we add to our dataset an independent variable that tracks the percentage of each ZIP code's population that is Norwegian. Now instead of regressing Y on X (ILLNESS on MAYOR VISITS), we now regress Y on X_1 and X_2 (ILLNESS on VISITS and %NORWEGIANS). We get the results just like before. And the equation reads like this:

ILL = 2.0 (Visits) + 5.8 (%Norwegians) + 106.

We can interpret this equation in the following manner: If we hold constant the percentage of population which is Norwegian, then for every one additional visit to the mayor, there will be an average of 2.0 more sick people. If we control for the number of mayor visits, then for each additional 1% increase in Norwegian population, there will be an average of 5.8 more sick people. Finally, if no one visits the mayor and there are no Norwegians, then we should still expect an average of 106 cases of illness.

VII. Example: Presidential Elections and Economic Variables

Now we can finally read a regression table, such as Table 3 (below), which is an example of regression results found in typical social science papers. Let's start with some background regarding the scientific debate to which Table 3 contributes. There is a disagreement amongst election campaigners and researchers over the degree to which economic conditions influence U.S. presidential elections. Some scholars argue that voters are

Table 3: Regression Analysis of Data on Presidential Elections and Economic Variables

	Regression 1	Regression 2	Regression 3	Regression 4	Regression 5
% Change in per capita GNP	0.63 (4.01)***	0.39 (2.57)**	0.63 (4.40)***	0.44 (3.31)***	0.41 (2.67)**
Change in Prices		-0.99 (-4.14)***		-0.78 (-3.52)***	-0.95 (-3.78)***
Inflation Rate (%)	-0.93 (-3.63)***		-0.66 (-2.62)**		
Previous Midterm Vote			0.57 (2.64)**	0.58 (3.00)***	
Previous Presidential Vote					0.15 (0.83)
Constant	53.7 (47.4)***	55.2 (42.8)***	25.0 (2.28)**	25.8 (2.61)**	46.5 (4.41)***
R^2	0.47	0.52	0.58	0.64	0.53
Observations	32	32	32	32	32

Notes: Dependent variable = the percentage of the two-party vote in the presidential election received by the candidate of the incumbent party. T-ratios in parenthesis.
*significant at the 0.1 level; ** significant at the 0.5 level; *** significant at the 0.01 level.
Adapted from: Lynch, G. Patrick (Dec 1999). "Presidential Elections and the Economy 1872 to 1996: The Times They Are a 'Changin or the Song Remains the Same?" *Political Research Quarterly* (52)4: 833.

fairly good at connecting specific government policies with their economic outcomes, and therefore tend not to blame Presidents for economic problems that are not directly their fault.[9] Others argue that voters hold the executive branch accountable for economic outcomes, even when the President is not directly responsible.[10] In order to resolve this debate, a researcher, Patrick Lynch, performed regression analysis on thirty-two

[9] Kramer, Gerald. 1983. "The Ecological Fallacy Revisited: Aggregate versus Individual-Level Findings on Economics and Elections, and Sociotropic Voting" *American Political Science Review* 77:92-107.
[10] Ritter, Gretchen. 1997. *Goldbugs and Greenbacks: The Antimonopoly Tradition and the Politics of Finance in America, 1865-1896.* New York: Cambridge University Press.

Presidential elections from 1872-1996 and found that both sides are partially correct: voters have historically held Presidents responsible for economic outcomes, but did so even more strongly after the executive branch gained greater power over economic policy during the Great Depression. Table 3 is part of Lynch's evidence, but how do we read it?

First, we find the names of the independent variables listed down the first column. The other columns each present a single set of regression results, each based on a different regression model (i.e. a different combination of X's). The notes below the table tell us that the dependent variable is the percentage of the two-party vote in the presidential election received by the candidate of the incumbent party. T-ratios are reported in the parentheses below each coefficient estimate.

Regression 1 tells us that a 1% increase in per capita GNP during the year of the election translates into an average increase of 0.63 in percentage of the two-party vote received by the incumbent. It also tells us that a 1% increase in inflation corresponds to an average decrease of around 1% of the two-party vote received by the incumbent. And the constant tells us that if there was zero economic growth and total price stability, then the incumbent would receive 53.7% of the two-party vote.

In Regression 2, Lynch substitutes "change in prices"[11] for "inflation" as his measure of price stability. Note that the slope coefficient does not change much. This use of different data to measure the same variable is called triangulation. And when triangulation produces similar results, it increases our confidence in those results. For example, if "change in prices" had a significantly different slope than "inflation", then we might doubt a general hypothesis about the effects of price stability on voting. In this case, it would suggest that we might need different hypotheses for inflation and deflation. Lynch also argues that the increase in R^2 from 0.47 to 0.52 suggests that "change in prices" is a better measure than "inflation" since that regression explains more of the data.

US presidential elections are usually fairly close; small percentage changes in voting can matter a lot. Therefore Regressions 1 and 2 provide evidence that economic growth and price stability do help, albeit in a small way, to shift US presidential election outcomes. Indeed, the R^2 of Regression 1 tells us that per capita GNP and inflation alone explain 47% of the variation in the percentage votes received by the incumbent.

Of course, the electoral fortunes of political candidates often depend on past successes. For example, you could argue that George W. Bush's strong performance in the 2004 election was partly built upon his victory in 2000, and upon Republican success during the 2002 congressional midterm elections. Therefore Lynch controls for these factors in Regressions 3-5. Here he shows that, even when you control for previous midterm or presidential election performance, economic conditions still matter. The coefficients for GNP per capita and the price stability variables do change somewhat, but they remain in the same general ballpark. Moreover, they remain statistically significant and their signs (i.e. +/-) do not change. We therefore retain our confidence in them as significant explanatory variables.

In his research paper, Lynch then runs more regressions (not shown here) in which he either includes dummies (see below) for individual election years, or splits the dataset into different time periods. In these regressions, the coefficients for GNP per capita grow larger, while those for price stability shrink, for later time periods. Thus Lynch concludes that, as Presidents gained more power to regulate economic performance, voters responded by holding them more accountable, especially for economic growth. Has he *proved* this hypothesis? Of course not. But he has produced some good evidence to support it. Future researchers can then use different data to support or contradict him. This is how scientific debate in all fields proceeds.

VIII. Data-Mining

One of the first thoughts to occur to students when they first learn multivariate regressions is: "Fantastic! I can just enter data for a bunch of variables, and then run regressions until I find something." This is called "data-mining", and although it might be useful for *generating* hypotheses in some fields, unfortunately it is not a scientific way to *test* a hypothesis in social science. In fact, data-mining is tremendously frowned upon in the social sciences because, without any prior hypotheses to test, it is never quite clear what you find when you data-mine.

For example, let's say you want to explain and predict the stock market. So as your Y, you enter in monthly data for the Dow Jones Industrial Average (DJIA) between 1990 and 2000. For your "X's", you then

[11] Where "change in prices" includes both deflation and inflation as positive numbers.

enter data on a bunch of random variables to analyze. One of the variables you dump in is data on the height of Heisman Trophy winner Jason White, who was in grammar school and high school during that time. Clearly, your regression results would suggest a positive relationship between the DJIA and White's height, because both grew significantly between 1990 and 2000. However, we know that they had nothing do with one another: White's growth did not cause the stock market to go up. The results of that data-mining process would be spurious.

You might argue that this is a ridiculous example. Someone would have to be completely blind to make these mistakes. But that's the whole point! If we blindly perform regressions on random data, then we will find all sorts of spurious results. Theory and hypotheses must come first! We then design a regression model that can best test our hypothesis. And only then do we gather data and run regressions.

Furthermore, there is the problem of significance. Remember the 95% confidence level that is traditionally used to signify a statistically significant finding. This 95% confidence level implies that, for any given independent variable, there is a 5% chance that it will be found to be "significant", when it really is not. This means that if we have 14 independent variables in our regression model, with each one having a 5% probability of being a fluke, then there is a 50% chance that at least one of them will be found to be "significant" merely as a result of our confidence level.[12] In other words, if we increase the number of independent variables in a regression, we increase the likelihood that one of them will be found "significant" merely due to chance rather than true correlation.

IX. Gauss-Markov Assumptions

Now that you understand some of the basic concepts of regression procedures and results, let's return to the Gauss-Markov Theorem. This theorem is important because it proves that OLS produces the best linear unbiased estimators (i.e. the best fitting lines). But it only works under certain conditions. Indeed, we have already seen examples of some of the things that can go wrong when using OLS. These are really just examples of Gauss-Markov assumptions that need to be fulfilled in order for OLS to work properly.

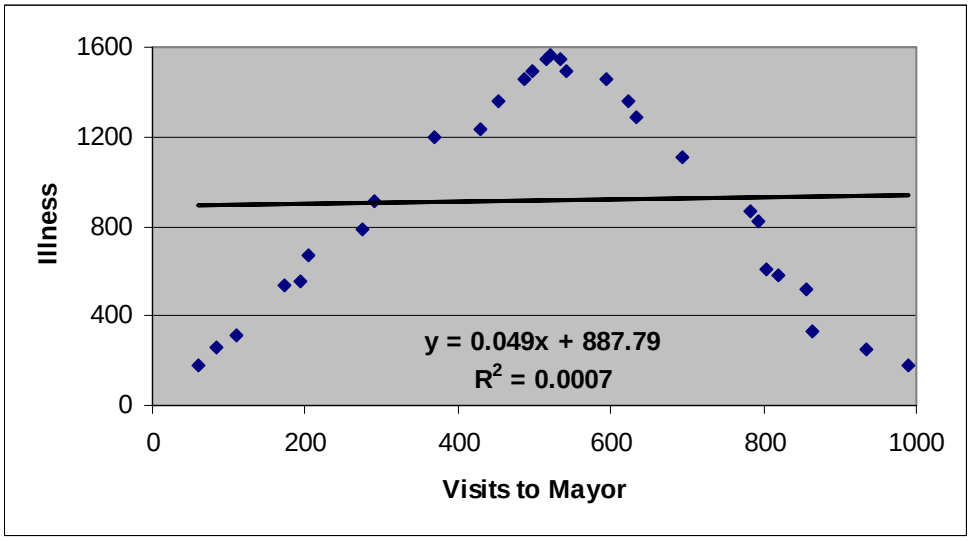
Assumption #1: A continuous dependent variable. OLS assumes that your "Y" is a continuous variable (e.g. population, gross domestic product, percentage of votes received). OLS does not work when your dependent variable is a category (Republican/Democrat/Independent, Socialism/Capitalism/Communism) or dichotomous (e.g. war/peace, win/lose, yes/no). In these cases, the fix is to use a slightly more advanced regression technique, such as probit, logit, etc. But there is no need to worry if any of the *independent* variables are a non-continuous variable; OLS can handle that just fine (more on this below).

Assumption #2: A linear relationship. Since OLS draws lines through datapoints, the relationship you hypothesize must be linear. The flipside to this is, just because OLS fails to produce significant coefficients, this does not mean that there is no relationship between X and Y. Look at the inverse-"U" shaped scatterplot below (Figure 6). There is clearly a relationship here between X and Y, but if you performed regression analysis on this data, the result would be a line with zero slope and a very low R^2. You would likely walk away from such a regression mistakenly thinking "no relationship here!". OLS would likewise fail to properly recognize exponential relationships, logarithmic relationships, quadratic relationships, etc.

What is the proper fix? Transform the data. That is, if you suspect a non-linear relationship, then perform a mathematical operation on the data that would turn it linear for the purposes of testing. For example, if we suspected the inverse-"U" shape below, we might divide the data and invert one half of it, and then use OLS to try to fit a line to it; or we might perform one regression on the lower half of the data (looking for a positive slope) and another regression on the higher half (looking for a negative slope). You can likewise use logarithms, exponents, and squares to transform other types of data where appropriate.

[12] $1-(0.95)^{14} = 0.49$, if this does not make sense, then review the multiplication rule and Chevalier de Mere's dice-game problem from introductory statistics.

Figure 6: OLS Attempts to Fit a Line to a Non-Linear Relationship



Assumption #3: The data are accurate and the sample is random. As with all statistical analysis, the results from OLS are only as good as the data. Hence OLS assumes that measurement errors are minimal, and that what errors exist are random. In other words, there should be no systematic bias in the data. So in the mayor-illness regressions above, if you had accidentally taken most of your data from heavily Norwegian neighborhoods (or from neighborhoods with no Norwegians at all), then you would get inaccurate regression results because your sample was not representative.

Beyond these three fairly obvious assumptions above, the way to think about the Gauss-Markov assumptions is to ask what can go wrong with your regressions. Since we care mostly about estimating the slopes correctly, then there are usually only two things that can go wrong in OLS: either the estimates of the coefficients can be off ("biased"), or the standard errors can be off ("inefficient"). Therefore we should focus on conditions that can cause these problems.

Assumption #4: No model specification error. Put simply, in addition to linearity (Assumption #2 above), this means that all relevant X's should be included in the model, and irrelevant X's should not be included in the model. In other words, it helps if you have the correct model. Omitting a relevant variable ("omitted variable bias") can result in biased estimation of the coefficients, while including irrelevant variables can inflate the standard errors of the other X's.

Why? Because if you control for an X that does not really matter, but is highly correlated with an X that does, it will steal some of its explanatory power. For example, say a lot of the people who were visiting the mayor and getting sick happened to be Democrats. If you included party affiliation (e.g. Democrat vs. Republican) as one of your regressors, then OLS would look at the data and say "wow, visits to the mayor matters a lot...and so does being a Democrat". However, we know that being a Democrat does not make you ill (regardless of how it might make Republicans feel). But if you were to include party affiliation in the regression model, then the coefficient for VISITS will be smaller than it should be because the coefficient for DEMOCRAT will steal from it.

Assumption #5: Homoskedastic errors. Homoskedasticity is Greek for "equally spread out-ness". It refers to the fact that OLS requires that the errors all have the same spread (variance) for each value of the independent variable. In our flu example above, the VISITS and NORWEGIANS data might "err" *differently* for different ZIP codes. The variance might be quite wide in downtown ZIP codes, where many people visit City Hall daily and others not at all. Meanwhile, out in the suburbs, people might generally visit the mayor once a year or less;

thus when viewed as a group, their individual visits are each closer to the group's mean. But OLS assumes homoskedasticity. If you have heteroskedastic errors, then OLS will still produce good coefficients, but the standard errors estimates will be too small. Therefore you could wind up mistaking a significant finding for an insignificant one. This is commonly a problem with regressions on involving data on multiple geographic areas (countries, states, cities) or organizations (firms, political parties). The solution is to use a slightly modified form of OLS which weights the squares it estimates. In practice, this usually just means entering an additional command in your computer software. When in doubt, you will want to do this since it creates a higher bar for standard error calculations, and therefore for significance.

Assumption #6: Errors are normally distributed. Ideally, if you could measure all of the residuals, and then plot them on a graph, they should have a Gaussian or "normal" distribution. This is possibly the least important assumption, but some researchers argue that where it holds true, then OLS produces the best estimates.

Assumption #7: No autocorrelation. This means that the residuals should not be correlated with each other across observations. This is rarely a problem with cross-sectional regressions. Cross-sectional regressions are those that analyze different units (e.g. nations, states, companies) during a snapshot in time. For example, a regression of economic growth in 100 countries during 2005 is a cross-section. However, if you want to analyze data across time (e.g. economic growth in the US from 1900-2008), known as "time-series", then autocorrelation becomes a problem. Why? Consider presidential popularity, economic growth, or the stock market. The value of these variables today depends, at least somewhat, on their value yesterday, and the day before, and the day before that. Therefore some of today's errors (wandering data), will be correlated with or caused by yesterday's errors, and the day before that, etc. This correlation implies that some other causal variable has been left out of the regression model. OLS does not deal with this well. There are techniques to diagnose autocorrelation, and there are some simple fixes available (such as including year dummies or a time trend variable). But you often have to use a different regression technique, such as time-series analysis (for single units observed over a long period of time, e.g. stock prices) or time-series cross-section (for multiple units observed over a period of time, e.g. cross-national comparisons of economic growth).

Assumption #8: The errors should not correlate with any of the X's. Some statisticians argue that this is the only important assumption. It is actually another way of saying that you have not left any variables out of your equation. How can the error estimates be correlated with any of the X's? Well, the error term actually represents all causal factors *not* included as individual X's. In our illness example above, this would include everything from random factors like the weather and nose picking, to possibly important factors like the number of Norwegians. And we know that if you omit a variable *and* it is correlated with any of the X's, then the regression will produce biased results.

X. Dummy Variables

Although OLS does not work when the *dependent* variable is dichotomous or categorical, it can handle dichotomous *independent* variables just fine. A dichotomous or "dummy" variable is a variable that can be coded only as "1" or "0". For example, you might use dummies to code variables like political party, ethnicity, country, war, etc. But the interpretation of dummy variables is very different from continuous variables. Specifically, you do not interpret the coefficients of dummy variables as slopes of a line. Instead, you interpret dummies as creating a separate regression line, with the same slope but a different intercept.

Let's see a simple example. Say you want to explain differences in worker salaries (Y = salary). You hypothesize that salary is a function of education. You also suspect that gender discrimination affects salary decisions, therefore you want to control for gender too. So your model is SALARY = SCHOOLING + GENDER. Salary and years of schooling are continuous numbers, but how should you handle gender? The solution is to create a dummy, say MALE, in which you code "1" for men, "0" for women. Note that you do *not* create a GENDER dummy since it would not be intuitively clear what "1" or "0" gender would mean. You also do not create *both* a MALE and FEMALE dummy since this would be redundant (knowing the value for MALE makes a FEMALE dummy unnecessary); this would also crash the mathematical solution to the regression, but we will leave that part of the explanation to your statistics class.

Next you would collect salary, schooling, and gender data on a large number of individuals. You would

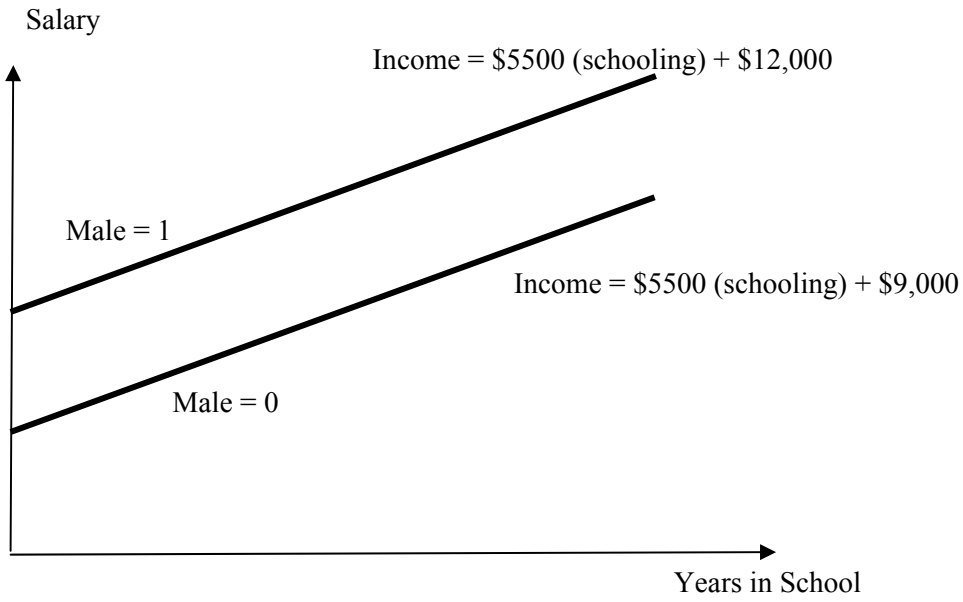
then run a regression on the data, and observe the coefficients and standard errors just like any other regression. But the interpretation of these coefficients is a bit different for the MALE dummy. Let's say the results produce a regression line like this:

Salary = $5500 (Schooling) + $3000 (MALE) + $9000

These results suggest that workers make $5500 more dollars in income for each additional year of schooling they receive. It also shows that males make $3000 more than females. In other words, the MALE dummy can be 1 or 0, while its coefficient is $3000. Together they multiply to be either 0 (for females) or $3000 (for males), hence dummies simply add to the intercept term. In other words, the model implies two different regression lines, one for females and one for males. The slopes are the same, but the intercepts are different, with the difference between the two lines being the dummy's coefficient.

Male Salaries = $5500 (Schooling) + $12000
Female Salaries = $5500 (Schooling) + $ 9000

Figure 7: Dummy Variables



Another way of interpreting this regression is that, since you always include one less dummy than the number of categories, then the coefficients for the dummies tell us how much effect the variable has relative to the missing category. This can be seen more clearly if we add race to the above equation. Race is a categorical variable, but not a dummy variable because there are more than two race categories. We therefore turn race into dummy variables by creating a dummy for each category we want to analyze:

Income = schooling + FEMALE + ASIAN + BLACK + LATINO

In this case, we suspect that racial discrimination also affects salaries. Let's assume that our hypothesis focuses us on four major race categories (Asian, Black, Latino, White). To control for race we would include in the regression model dummies for only *three* of these categories, say Asian, Black, and Latino. But we do *not* put in a dummy for WHITE because it would be redundant. The resulting coefficients for Asian, Black, and Latino would therefore tell us how much workers in these race categories earn relative to workers in the missing category, White. Let's say the regression results look like this:

Salary = 3723 (Years in school) + 509 (Male) -680 (Asian) + 1920 (Black) -900 (Latino) + 15680

These hypothetical regression results suggest that, at least for the sample for which we collected data:

1) Each year of schooling results in a salary increase of $3732 (holding race and gender constant).
2) Males make $509 more than females (holding race and schooling constant).
3) Asians make $680 less than whites (holding gender and schooling constant).
4) Blacks make $1920 more than whites (holding gender and schooling constant).
5) Latinos make $900 less than whites (holding gender and schooling constant).
6) The salary of an uneducated, white, female (i.e. a value of "0" for all variables) is $15680.

XI. Interaction Terms

A final device often used in social science regressions are "interaction" or "multiplicative" terms. An interaction term multiplies two independent variables together. They are used to model hypotheses in which the effect of X_1 on Y is conditional on X_2 (or vice-versa). Let's see how this might work.

Sticking with our example above, say we hypothesize that salaries are a function of education and experience. But we believe that the effects of education on salary are conditional on experience. That is, the effect on salary of a worker's education will depend on her experience on the job. An experienced worker will get a larger salary bump from an MBA than someone fresh out of college. Such a regression model would look like:

$$\text{Salary} = X_1\,(\text{education}) + X_2\,(\text{experience}) + X_3\,(\text{education*experience})$$

Where (education*experience) is the interaction term.

The interaction term's coefficient X_3 tells us the effect on income of being educated and experienced not explained by education and experience considered separately. More precisely, X_3 tells you how much the effect of education (on salary) changes per unit increase in experience (and vice versa). It therefore tells you how the effect of education (on salary) is conditional on experience (and vice versa). I keep repeating "vice versa" here because the two statements are mathematically equivalent. Statistics cannot tell you which independent variable drives the other in the interaction term. Rather, good theory should come first, and inform us how to interpret them.

Notice that X_1 no longer tells you the *average* effect of a unit increase of schooling on income, holding experience (X_2) constant! Instead, the coefficient on education now tells you the effect of one year of education when experience = 0.

Likewise, X_2 no longer tells you the average effect of a unit increase of experience on income, holding education (X_1) constant! The coefficient on experience now tells us the effect of one year of experience when education = 0. This can be seen a lot easier in an example. Say we run the above regression on salary, education, and experience data and produce the following coefficients:

$$\text{Salary} = 3000(\text{education}) + 700(\text{experience}) + 250(\text{eductn*exprnce}) + 5000$$

Therefore, if we start with a totally uneducated worker, and then add education one year at a time, we would begin to get the results in Table 4:

Table 4

when education = 0						
Salary =	0	+ 700(experience)	+	0	+ 5000	
when education = 1						
Salary =	3000	+ 700(experience)	+ 250(experience)		+ 5000	
when education = 2						
Salary =	6000	+ 700(experience)	+ 500(experience)		+ 5000	

The regression results tell us that an uneducated worker with no experience could expect a salary of $5000.

Every year of additional education would add $3000 in salary directly from schooling, but also some additional income that would depend on the amount of experience (i.e. interaction term). Hence the "3000" coefficient tells us the effect of one year of education when experience = 0. The "250" coefficient tells us how much the effect of education (on income) changes per unit increase in experience (or vice versa). Note that this means that the effect of education on income is different for different levels of experience (and vice versa)! Workers with only one year of education, can expect each year of experience to add $700 + $250 to their income. But workers with two years of education will get more out of their experience: $700 + $500. Therefore the effects of education are conditional on the amount of experience (and vice versa).

When including an interaction term, many researchers also include its components separately, as we did above. This is because they want to show that the interaction term is significant even after they control for its components. However, there is nothing wrong with a theory that hypothesizes that the interaction term alone is what matters.

XII. What's Next?

The English poet Alexander Pope once wrote that "a little learning is a dangerous thing". So consider yourself warned: you have now learned enough about regressions to be dangerous. But in order to be useful, you need to learn more. The good news is that if you have taken the time to understand and memorize the basic concepts described in this paper, then learning more will be easy. In fact, Sage Publishing offers a special series, "Quantitative Applications in the Social Sciences", consisting of over 160 little green booklets each dedicated to different statistics topics. If you mostly understood this paper, but want to nail down some individual concepts better, or dig a bit deeper, then these booklets should be your next step. They are generally very well-written, highly accessible, light on math and theory, but heavy on the examples and applications. Some of the most relevant booklets have been listed below, along with some very useful textbooks, articles, and book chapters.

Recommended Readings

Berry, W. D. (2000). *Understanding Multivariate Research: A Primer for Beginning Social Scientists*. Westview Press.
If you didn't completely understand this paper, or you need a good basic repeat of the concepts, then this is the book for you. It's a simple, easy introduction to regression analysis.

-Greene, W. H. *Econometric Analysis* (Prentice Hall)
If this paper was child's play to you, then you might want this thick textbook. It's a widely considered to be "the Bible" of regression analysis. WARNING: its not very accessible to new students since it's mostly written in mathematical Greek. But if you're a math-jock, then have at it.

-Williams, F. *Reasoning With Statistics: How to Read Quantitative Research* (Wadsworth)
For those who simply want to consume, but not necessarily review or produce, quantitative research...recommended by a methods professor.

The little green Sage monographs

If you mostly understood this paper, but want to nail down some individual concepts, or dig a bit deeper, then these booklets should be your next step. They are absolutely fantastic, and I highly recommend them! Its a series of about 130 little booklets, each one tackles a different aspect of statistics. They are generally (though not always) very well-written, highly accessible, and are light on the math and theory, but heavy on the examples and applications. (OK, yeah, once in a while a poorly written one gets published). They are like little "how-to" books for statistics. I highly recommend the following...

General concepts and applications of OLS:

Lewis-Beck, M. S. (1980). *Applied Regression: An Introduction.* Thousand Oaks, CA: Sage.
Achen, C. H. (1982). *Interpreting and Using Regression.* Thousand Oaks, CA: Sage.
Berry, W.D., & Feldman, S. (1985). *Multiple Regression in Practice.* Thousand Oaks, CA: Sage.
Schroeder, L, Sjoquist, D.L., & Stephan, P.E. (1986). *Understanding Regression Analysis: An Introductory Guide.* Thousand Oaks, CA: Sage.

Berry, W.D. (1993). *Understanding Regression Assumptions.* Thousand Oaks, CA: Sage.

Special topics in OLS:
Fox, J. (1991). *Regression Diagnostics: An Introduction.* Thousand Oaks, CA: Sage.
Hardy, M.A. (1993). *Regression with Dummy Variables.* Thousand Oaks, CA: Sage.
Jaccard, J.& Turrisi, R. (2003). *Interaction Effects in Multiple Regression.* Thousand Oaks, CA: Sage.

Other Good Regressions Textbooks:
Wooldridge, J.M. (2008). *Introductory Econometrics: A Modern Approach.* Mason, OH: South-Western.
Another great textbook, a bit more user-friendly than Greene, with a more examples and applications. It's new, but it's already widely assigned amongst many top econometricians.

Gujarati, D.N. (2009). *Basic Econometrics.* New York, NY: McGraw Hill.
A textbook, but this time written mostly in plain English. Much more accessible than either Greene or Wooldridge but not as deep & thorough.

Kennedy, Peter *A Guide to Econometrics* (MIT Press)
kind of an odd set of side notes explaining various regression concepts...some love it, some don't

Classic Articles on Regressions
Chatterjee & Wiseman. (1983). Use of Regression Diagnostics in Political Science Research. *American Journal of Political Science* 27, 601-613
King, G. (1986). How Not To Lie With Statistics: Avoiding Common Mistakes in Quantitative Political Science. *American Journal of Political Science* 30, 666-687.
Achen, C. (1990). What Does 'Explained Variance' Explain? *Political Analysis* 2, 173-184
Lewis-Beck, M. S. & Skalaban, A. (1990). The R-Squared: Some Straight Talk. *Political Analysis* 2, 153-171.
Brambor, T. & Clark, W.R. (2006). Understanding Interaction Models: Improving Empirical Analyses. *Political Analysis* 14, 63-82.

Popular Statistics Software Books
Philip H. Pollock, P. H. (2006). *A Stata Companion to Political Analysis.* Washington DC: CQ Press.
Philip H. Pollock, P. H. (2008). *An SPSS Companion to Political Analysis.* Washington DC: CQ Press.

Appendix I: How To Read Log v. Linear Terms in Regressions[13]

What's a logarithm?

1. Put simply, if $b^y = x$, then $\log_b(x) = y$.

2. More specifically, the logarithm of a number (x) to a base (b) equals the power or exponent (y) to which the base must be raised in order to produce the number.

3. Example A: **$\log_{10}(1000) = 3$**
   You would read this as: the logarithm of 1000 to the base 10 is equal to 3. The reason "3" is the answer is that $10^3 = 1000$.

4. Example B (using a different base): **$\log_2(32) = 5$**
   This is true because $2^5 = 32$.

5. Finally, note that the logarithm of x to the base b is written $\log_b(x)$ or, if the base is implicit, as log(x).

What's a "natural logarithm"
If the base of the logarithm is chosen to be Euler's number or "e" (where e = 2.718), that is, using **$\log_e$** , then it is denoted by "ln" as in ln(x) = y and called a "natural logarithm".

Why do we care about logarithms?

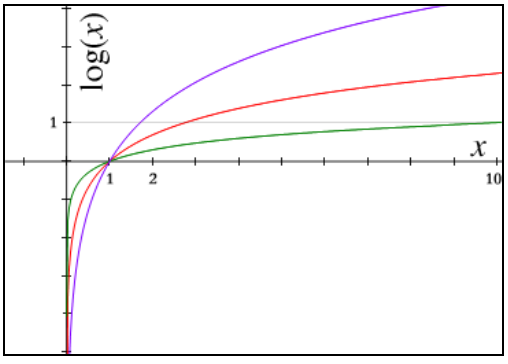
1. Many variables, especially in economics, are generally thought to have a log relationship with one another rather than a linear relationship.

   Example: how do population and the economy relate to one another? Many economists argue that **GDP = β * ln(population)** is more accurate than **GDP = β * (population)**, where β is just some coefficient.

2. Logs are also good at bringing outlier datapoints in from the extreme.

   Example: if your data ranges from 10 to 1,000,000 then you can log the data to get a tighter range of 2 to 14. This will prevent outliers from skewing the regression results.

3. We can see a graph of various $\log_b(x) = y$ relationships below, with the different colored lines corresponding to different types of log (base e, base 10, base 1.7).



The important thing to note in this graph is that log relationships are *not* linear. And if they are not linear, then we cannot use OLS to test for them (since OLS fits *straight lines* to data). The solution then is to "log" the data. In this manner, a log relationship will become linear, thus allowing OLS to fit a line to it.

[13] Also known as "log-linear models", not to be confused with "logistic regression" which is an altogether separate subject.

So What's the Problem?

The problem then becomes: how do we interpret the coefficients that OLS gives us? The answer is: percentages.[14]

Example: say that an analyst believes that the **price of a home** has a relationship with the **size of the property lot** it sits on. Therefore she gathers data on lot sizes and housing prices, and runs some OLS regressions on them. Take a look at the example results below to see how she should interpret different types of log relationships. Note that in some regressions below, the dependent variable is logged, in others the independent variable is logged. (The regression results are shown in bold, the interpretation is below in plain font).

==========LOG-LOG==========

ln(PRICE) = 0.35 ln(LOT SIZE) + constant

A proportional increase in X corresponds with a proportional increase in Y, therefore:
A 1% increase in LOT SIZE increases the PRICE by 0.35%

==========LINEAR-LOG==========

PRICE = 19758 ln(LOT SIZE) + constant

A proportional increase in X corresponds with an absolute (or unit) change in Y, therefore:
A 1% increase in LOT SIZE increases the PRICE by 197.58 (= 19758 * 0.01)

==========LOG-LINEAR==========

ln (PRICE) = 0.00008 LOT SIZE + constant

A unit increase in X corresponds with a proportional change in Y, therefore:
A 1 unit increase in LOT SIZE causes a 0.008% (= 100 * 0.00008) change in PRICE.

==========LOG-DUMMY==========

ln (PRICE) = 0.136 DRIVEWAY + constant

The difference in ln(PRICE) between a house with a DRIVEWAY and one without a DRIVEWAY is 0.136. So a house with a driveway is about 13.6% more expensive than one without. In general, however, we would calculate the exact percentage difference from $100(\exp^{coeff} - 1)$.

[14] Since they are read as percentages, log-log regression models are often used to compute elasticities in economic analysis (the ratio of the percentage change in one variable to the percentage change in another variable).