

1 Linkage of Theory to Data

What does it mean to link theory to data? It means you have an expectation. You have a view of the way the world works (or that part of the world that you're interested in). What constitutes a "real" theory, is, of course the stuff philosophy of science courses are made of. (Popper, Kuhn, Lakatos).

Additionally, there is the literature on grand theories versus middling theories ... the point is, there is some prior expectation out there on how you think the world behaves. Rational choice theorists posit, that political action is born out of self-interest. Self-interest then leads to some *observable implications*.

For example, in a legislative setting, institutional arrangements may be established to promote reelection. If one adopts a rational choice perspective—the theory—drives one's expectations.

In the late 19th Century, most U.S. states had adopted the Australian Ballot. The AB placed all candidates on a single ballot. With the onset of the AB, we started to observe split-ticket voting more-and-more often. Rational politicians had to adapt. No longer could they simply rely solely on party labels.

A (simple) rendering of a rat. choice perspective would lead us to expect changes in politician's behavior. These expectations are our observable implications. One observable implication might be that the pursuit of the "personal vote" will emerge.

That is, politicians, now under the condition that they're more electorally responsible for their careers, may adjust behavior in order to heighten the probability of reelection. An observable implication might involve committee tenure. The longer one stays on a congressional committee, the better the ability the member has in procuring "pork" for the district or simply representing the district.

Therefore, tenure on a congressional committee might increase after the onset of the AB. This observable implication is a testable hypothesis. And what's more, the hypothesis is embedded (albeit simplistically) in a theory of rational behavior.

The test of this hypothesis would require data and analysis. But the crux of the test emanates *not* from statistics, but from theory. Thus, it doesn't really make any difference if you're doing quantitative or qualitative work: the endeavor is exactly the same.

What is the alternative to theory-based research? Search and destroy missions? Exploratory analysis? Is there anything wrong with this? In and of itself, no. The problem is, people are often dishonest in their use of exploratory analysis. You can run into the problem of "disconfirmability."

2 Why Quantitative Methods?

My take on why we use quantitative methods revolves loosely around generalization and I guess in some sense replicability (although either of these things are satisfiable in a qualitative setting if the design is thorough). The methods we consider have structure and have assumptions. They are grounded in mathematical and probability theory. The methods hang by their assumptions and if the assumptions are wrong given the data, then it makes it possible for others to come along and replicate or not replicate findings.

A frequently heard complaint against quantitative methods is that they are too complex or too rigid for an inherently complex and ever-changing world.

To some extent, I sympathize. Statistical methods are often routinely used incorrectly. Assumptions are made that quite simply cannot be sustained by the data.

But far from giving up the ship—or as some might proceed—forgoing the “science” of politics—I think the complexity of the world demands a fairly complex understanding of methodology. Or to put it another way, it is difficult to understand complex phenomenon with grotesquely simplistic methods (whatever those methods may be).

Quantitative methods provides a tool for understanding the world.

3 Scope of the Method

Although we often hear reference to phrases like “stats”, what we are really aiming for here is to develop a sense of making inferential statistics.

But inferential statistics has the aim of making some generalization to a **population** of interest. What is the population? It is research determined.

Problem is, we seldom have available a population, so we rely on sampling data. As soon as we enter the realm of sample-based data, then we immediately are making strong assumptions about the nature of data collection.

Were the data collected independently? Were the data collected randomly? What was the sampling frame? Answers to these questions bear on the inferences made.

Much of our research, however, is based on data not generated from a sample. For example, we look at aggregate data frequently in international relations or in comparative politics. Frequently, you will hear someone say that statistical tests of significance need not be done because a full population of data are available.

Clearly, the idea here is that since inference is based on the notion of generalizing to a population, if you have a population, you don’t need to generalize.

This view is mythology. The data generation process has randomness built into it. Although you may have a “population”, there is considerable amounts of measurement error because our political concepts are too complex to ever get “right.” Even with apparent populations, we are still making inferences.

3.1 On Modeling and Explaining the Variance

The notion of inference is then rooted squarely in probability theory. All of our methods that we learn in here have assumptions that are grounded in fairly straightforward “rules” of probability. If the probabilistic assumptions can’t hold, it becomes exceedingly difficult to make inferences.

Eventually, we will get to a point of modeling. We will treat some factor—an endogenous factor—as a function of something theoretically interesting—exogenous factors:

$$Y = F(X) \tag{1}$$

The hard part will be specifying $F(X)$. And indeed, we won’t learn very many sophisticated ways of specifying $F(X)$. Nevertheless, what we need to understand is that there is *unaccounted-for* variance in the world:

$$Y = F(X) + e \tag{2}$$

Although when we learn regression, we’ll learn very technical details about the nature of “e”, for now, understand that “e” corresponds to error—to random error. Another way of looking at

it, is that “e” corresponds to all the other influences in the world, unaccounted for in our model” that exert pressure on Y.

Can we perfectly account for “e”? Probably not and we probably do not want to. The more complex a model becomes, the more difficult it becomes to interpret a model—a road map where 1 mile equals 1 mile is not very useful to us.

But “e” is important when we talk about “explaining the variance.” What is the variance? If the world could be determined by model 1, we’d have a deterministic model. For example, if the number of political riots in a country was determined by income inequality (however measured), *and nothing else*, then we would have a deterministic model.

The problem is, there is wonderful randomness in the world that belies perfect predictability. Notationally, this is expressed as “e.” It is error or randomness we can’t account for or choose not to account for. It doesn’t go away but it causes variation to persist in our data. This is **stochasticity**.

Thus if we have an endogenous and exogenous variable, we are trying to see how much in the variation across Y our knowledge of X brings to bear in explaining the variance in Y.

This is the notion of the term “explaining the variance.”

4 Linkage of Analysis to Research Design

You can learn a lot from a standard error. Although you’ll learn about this later, I want to consider what the standard error looks like for a slope coefficient from a regression model.

Consider the following model:

$$y = \beta_0 + \beta x + \epsilon$$

Let’s call this a regression model where β represents the slope coefficient, which gives us an estimate of how y changes for some unit change in the value of x . (Recall the slope-intercept formula: $y = mx + b$.)

Because we deal with stochastic relationships—everything is probabilistic—we have to recognize there is error around our estimate of the slope coefficient. Omitting the mathematical details that gets us to this uncertainty, consider the following:

$$s_{\hat{\beta}_i} = \sqrt{\frac{\Sigma(y_j - \hat{y}_j)^2}{\Sigma(x_{ij} - \bar{x})^2(1 - R_i^2)(n - k - 1)}}$$

What do we have here?

First of all, in general, would you “prefer” a standard error to be larger or smaller? Smaller: Why? More precise inferences are permitted. Well, what makes this “thing” small. When the stuff in the denominator is big relative to the stuff in the numerator, then what happens? (The s.e. decreases). When the stuff in the numerator is big relative to the stuff in the denominator, the s.e. increases.

So what does this say about research design and quantitative methods?

First, what is the term in the numerator representing? (Technically, it is the sum of squares due to error). But it represents deviations from a model’s predictions and the observed values of y . Smaller deviations, better predictions.

But more importantly, for now, is the stuff in the denominator. After all, data are full of error (selecting on y should be noted as problem). What does the first term represent? It denotes the variance of x . What is the implication here? More variance on x , the better off you will be. Research design issues: collect more data on x !

Second term: known as auxiliary regression. Think of it, for now, as an indication of how highly correlated two independent variables are such that a high R^2 denotes high correlation. What does this say about x ? The extent to which your independent variables are *not* measuring unique concepts effects your ability to make inference. Relate to collinearity. Note that with perfect collinearity—two variables are a linear combination of one another—the s.e. is undefined.

Third term: cases. If n denotes the sample size, then what happens as n increases? S.E. decreases (an added benefit is that the variance in x is guaranteed to go up and normal approximation, via the central limit theorem) improves. What is k ? This term represents the number of independent variables (1 represents the constant term). What happens as $k \rightarrow n$? This term gets smaller and the s.e. gets bigger. Implication? Small samples coupled with a large number of covariates is a bad thing.

By the way, dropping out covariates is a bad thing too, if one has no theory for dropping them out! (More on this, later).

5 Conceptual Regression: Preliminaries

How do basic statistics get us to the intuition of linear regression? Let's go back to the Law of Large Numbers. What does the law of large numbers say?

Draw observations at random from any population with a finite mean μ . As the number of observations drawn increases, the mean \bar{x} of the observed values gets closer and closer to the mean μ of the population.

So what?

The logic here is that if you collect enough data (and collect it in a valid manner), then your statistic will *approach* the population parameter. This is a good thing, as it gives us confidence that we will converge to the right population parameter.

The Law of Large Numbers tells us something about estimating the parameter, but it doesn't tell us much about the nature of the parameter. Enter the notion of a sampling distribution.

What is a sampling distribution?

The sampling distribution of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

This is textbook. The importance, however, is immeasurable. Why? Understand what a sampling distribution implies. It implies that in repeated samples, it is very likely that the estimate of a parameter may vary. (Coin flip example; or talk about feeling thermometers in repeated samples of 1000). This means that for every statistic that you estimate, there will be uncertainty around that estimate. Why? BECAUSE the statistic comes out of some sampling distribution—a distribution that tells us that the value of the statistic is variable. (Feeling thermometers are a random variable: a real valued function (there exists a score) defined on a sample space (perhaps our sample of 1000)).

Well we have some important things here. Let's think of the mean, \bar{x} . We know that the mean of the statistic \bar{x} is always the same as the mean μ of the population. We know this from the law of

large numbers and through the concepts of the sampling distribution. If the mean of the sampling distribution of $\bar{x} = \mu$, then we know, because of sampling variability, that *any particular* realization of \bar{x} may not *exactly* equal μ . (Suppose true approval rating of a president was 55 percent; any particular sample might give us a number above or below that).

However, because data are assumed to be *randomly* generated, then we assume there is *no systematic tendency to overestimate or underestimate the parameter μ* . That is, there is no **bias**. This is the essence of the idea of unbiasedness that you learned in POL 582. An unbiased estimator is correct on average.

Now, understand that we have gotten pretty far in connecting our statistic to our parameter. But one crucial element is missing: the **shape** of the sampling distribution. One fundamental fact from statistics is that if the distribution in the population is normal, then the sampling distribution of a statistic is also normal.

But what happens when the population is nonnormal? This leads to one of the most important results in statistics. First, remember to keep in mind the idea of a sampling distribution. (What is it?) As the sample size increases, the *sampling distribution* of \bar{x} changes its shape. It looks less like that of the population and more like a normal distribution (remember: we're dealing with a sampling distribution). When a sample is large enough, the sampling distribution of our statistic, \bar{x} is *very close to normal*. This famous result is known as the **Central Limit Theorem**:

Draw a random sample of size n from any population whatsoever with mean μ and finite standard deviation σ (this means that the variance is not ever increasing). When n is large, the sampling distribution of the sample mean \bar{x} is approximately normal.

The Central Limit Theorem gets us to the normal distribution. This is nice because now we can rely on the mathematical properties of the normal to help us draw inferences (standardized scores, areas under the curve, and so forth).

Now, back to regression. In the simplest incarnation of a linear regression model, we're interested in accounting for variation in the dependent variable, usually referred to as y , as a function of some independent variable, usually referred to as X .

Notationally, we may say that y is some function of X , such that $y = f(x)$. Now what if x had *no* relationship whatsoever with y ? If we created a scatterplot of y versus x , what might we find?

DRAW GRAPH WITH CLOUD OF POINTS (i.e. 0 slope)

What if y was related to x . Then we might find some apparent systematic relationship in the scatterplot.

DRAW GRAPH WITH POSITIVE CLOUD

Here, it seems as if values of y are increasing as values of x increase. Therefore, the y are *conditional* on values of x , that is $y|(x_1, \dots, x_k)$. The question naturally arises as to what aspect of y is conditional on x ? The natural choice is the mean. Why the mean? The mean is a good descriptor of central tendency under what conditions?

When we have a normal distribution, the mean best describes the center.

6 Regression as a Conditional Expectation

So what *is* going on in regression? What we're doing is modeling the linear relationship between a dependent variable and some independent variable. Our goal? Prediction? Inference? Ultimately, we want to say something about how y varies as a function of x . The simplest way to do this is in the context of a linear model, which takes the form $y = a + bx$.

The question, naturally arises, as to where do the numbers come from in the linear model. That is, where do the slope and intercept come from.

We're going to mainly focus on the slope for now, because we're keeping this at a conceptual level, and because one is usually interested in the slope for making inferences and predictions.

Let's consider the money-votes data. (Go Thru Example, discussing the general relationships between money and votes).

As we see through this exercise, a natural way to describe the relationship between money and votes is to see how y conditional on some value of x varies as x increases. This is the basic intuition of regression. More formally stated, a regression model is a formal means of expressing the *tendency* of y to vary with x . More specifically, it is the relationship of the central tendency of y conditional on the x that we're interested in. The logic here is simple.

We assume there exists a probability distribution of y for every level of x . The means of these probability distributions vary in some systematic fashion with x . We're assuming the y is a random draw from each value of x (yet another reason why one shouldn't usually select on the dependent variable). Therefore, what a regression model is going to give you, is the *expected value of y for a given value of x* .

And as you recall from 582, the $E(Y) = \bar{y} = \mu$. In regression, we're modeling $E(Y)|x_1, \dots, x_k$, that is, the **conditional expectations**. This gives us the leverage to say something about the relationship between x and y . (Question: what if in the population $\mu|x = \mu$?). That is, what if the mean of the probability distribution of y was equal to the mean of the probability distribution of y given x ? We would have no relationship.

We're going to take advantage of this idea to develop the regression model further next time.

Of course the notion of conditional means gets us only so far. The mean is good so long as the conditional distribution is approximately normal. Skewness

Heavy Tails

Unequal Spread

Bimodality

Nonlinearity

7 Introduction to Regression (Bivariate Case)

Last time, we talked about the fact that in regression, we're really modeling a conditional relationship:

$$E(Y_i | X_i) = \beta_0 + \beta_1 X + e,$$

where the coefficients are estimate of the population parameters and e is a stochastic disturbance term. The e term is also known as a *residual* term. (What is a residual)? Leftovers. Why do we posit the existence of e ? Stochastic world; simplified models. What must we assume about e ? We assume that since the leftovers are random, there is no systematic variation in the term e .

With algebraic manipulations, what does e look like?

$$e_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

and so if

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

then the residual term is equal to

$$e_i = Y_i - \hat{Y}_i.$$

This is useful to us, for it tells us in an obvious way, that the residuals, the leftovers, are a function of the observed data and the predicted data. (DRAW PICTURE OF RESIDUAL). Ideally, we would prefer the residuals to be small, rather than large.

In some sense, “making” the residuals small is easy. After all, note that

$$\Sigma(e_i) = \Sigma(Y_i - \hat{Y}_i)$$

will *always* be equal to what number? (Zero) Why? Deviations above the predicted line cancel out deviations below the predicted line. The problem is that *any* line passing through the means of X and Y (which is guaranteed to occur in the regression setting) will yield a sum of the residuals equal to 0.

Still, our goal is to minimize the residuals:

$$\min \Sigma \hat{e}_i = \min \Sigma(Y_i - \hat{Y}_i).$$

The trick is, how to do this. This gets us to the least squares principle.

8 Extension from Bivariate Case

The extension of the bivariate linear regression model to the multiple regression setting is straightforward. Simplistically, it entails the inclusion of 2 or more additional covariates, x_i . The least squares principle extends to the n -variable setting in a straightforward way, and the interpretation learned regarding a single x extends directly to the multiple regression setting.

To fix ideas, let us define a multiple regression model in the following way:

$$Y_i = \beta_0 X + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i.$$

In contrast to the two-variable model, we have two slopes and an intercept to worry about. Note that the \hat{y} computed from a two-variable model will produce a straight line; the \hat{y} computed from the multiple regression model with two independent variables will produce a *surface*; and a multiple regression model with $n > 2$ covariates will produce a function that is hard to visualize. Yet all the variants of the linear regression model start with the least squares estimator. As such, the basic quantity of interest, from the standpoint of statistical estimation, is

$$\begin{aligned} \hat{e}_i &= Y_i - (\hat{\beta}_0 + \hat{\beta}_{11} X_{1i} + \hat{\beta}_2 X_{2i}) \\ &= Y_i - \hat{Y}_i, \end{aligned}$$

the residual term. Like two-variable regression, the least squares solution produces estimates of the parameters $\beta_0, \beta_1, \beta_2$ (that is $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$) that minimizes the SSE:

$$\begin{aligned} \min \Sigma \hat{e}_i^2 &= \Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})^2 \\ &= \Sigma(Y_i - \hat{Y}_i)^2. \end{aligned}$$

The basic idea (like before) is to find estimates of the parameters that minimizes the SSE. Mathematically, this entails differentiation of

$$\Sigma \hat{e}_i^2 = \Sigma(Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})^2$$

partially with respect to the three unknown parameter estimates. This gives

$$\begin{aligned} \frac{\partial \Sigma \hat{e}_i^2}{\partial \hat{\beta}_0} &= 2\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})(-1), \\ \frac{\partial \Sigma \hat{e}_i^2}{\partial \hat{\beta}_1} &= 2\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})(-X_{1i}), \\ \frac{\partial \Sigma \hat{e}_i^2}{\partial \hat{\beta}_2} &= 2\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})(-X_{2i}), \end{aligned}$$

which when set to 0 and rearranging terms produces the *normal equations*:

$$\begin{aligned} \bar{Y} &= \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2 \\ \Sigma Y_i X_{1i} &= \hat{\beta}_0 \Sigma X_{1i} + \hat{\beta}_1 \Sigma X_{1i}^2 + \hat{\beta}_2 \Sigma X_{1i} X_{2i} \\ \Sigma Y_i X_{2i} &= \hat{\beta}_0 \Sigma X_{2i} + \hat{\beta}_1 \Sigma X_{1i} X_{2i} + \hat{\beta}_2 \Sigma X_{2i}^2. \end{aligned}$$

These equations can be rewritten yet again in terms of the parameter estimates:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\Sigma(X_1 - \bar{X}_1)(Y_i - \bar{Y})(\Sigma(X_2 - \bar{X}_2)^2 - \Sigma(X_2 - \bar{X}_2)(Y_i - \bar{Y})(\Sigma(X_1 - \bar{X}_1)(X_2 - \bar{X}_2))}{\Sigma(X_1 - \bar{X}_1)^2 \Sigma(X_2 - \bar{X}_2)^2 - (\Sigma(X_1 - \bar{X}_1)(X_2 - \bar{X}_2))^2} \\ \hat{\beta}_2 &= \frac{\Sigma(X_2 - \bar{X}_2)(Y_i - \bar{Y})(\Sigma(X_1 - \bar{X}_1)^2 - \Sigma(X_1 - \bar{X}_1)(Y_i - \bar{Y})(\Sigma(X_1 - \bar{X}_1)(X_2 - \bar{X}_2))}{\Sigma(X_1 - \bar{X}_1)^2 \Sigma(X_2 - \bar{X}_2)^2 - (\Sigma(X_1 - \bar{X}_1)(X_2 - \bar{X}_2))^2} \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 \end{aligned}$$

Several things are worth noting. First, this system is identified because there are three equations and three unknowns. Under pretty general conditions, you *will* get estimates using the least squares approach. Indeed, the only conditions under which you will not get estimates are when X_1 and X_2 are perfectly correlated. To see this, note that if

$$\Sigma(X_1 - \bar{X})^2 \Sigma(X_2 - \bar{X})^2 = (\Sigma(X_1 - \bar{X})(X_2 - \bar{X}))^2,$$

then it must be the case that one variable is a linear function of another variable; that is, the two variables, under this condition, are perfectly *collinear*, thus producing a correlation coefficient of exactly 1 (or -1). Indeed, the above expression can be rewritten in terms of the correlation coefficient,

$$r_{X_1, X_2} = \frac{\Sigma(X_1 - \bar{X})(X_2 - \bar{X})}{\sqrt{\Sigma(X_1 - \bar{X})^2 \Sigma(X_2 - \bar{X})^2}},$$

making it easy to see the condition under which $r = 1$ (the product mean deviations of xy is equal to the product of the standard deviations of x and y . What would produce this? Any time X_1 is a perfect linear combination of X_2 (or vice versa), this will occur. Example: $X_2 = X_1 \times K$ (where K is a constant. Other examples where this can occur involves the use of binary variables (more on this later in the semester).

Another situation where nonestimation of the parameters will occur is when the standard deviation of one of the regressors is 0—that is, one of the X_k is constant. This is trivial: seldom

will you run into this problem. However, suppose that the variance on one of the covariates is very small. The implication is that there will be a limited amount of information upon which the coefficient estimate will be based (so even though you can estimate it, the coefficient will be highly unreliable).

To illustrate, let's work through some examples. I've created a data set and sent it to you called `simulateddata.dta`. I've also sent along the log file so you can replicate.

Let's consider the happy case where two independent variables have a low correlation. To do this, I created two independent variables using Stata's random number generator. The correlation between the two variables is .0002 (nearly 0, but not quite). Also, a dependent variable was simulated. First of all, let's use information regarding the variances and covariances to compute the regression coefficients *without using regress*.

The following quantities are necessary in order to solve for the regression estimates:

$$\begin{aligned}\bar{Y} &= .4918 \\ \bar{X}_1 &= .4969 \\ \bar{X}_2 &= .5095 \\ \Sigma(X_1 - \bar{X})^2 &= 8.7018 \\ \Sigma(X_2 - \bar{X})^2 &= 8.2850 \\ \Sigma(X_1 - \bar{X})(X_2 - \bar{X}) &= .0018 \\ \Sigma(X_1 - \bar{X})(Y - \bar{Y}) &= -.1432 \\ \Sigma(X_2 - \bar{X})(Y - \bar{Y}) &= 1.1040. \\ (\Sigma(X_1 - \bar{X})(X_2 - \bar{X}))^2 &= .0000033\end{aligned}$$

First, let's compute $\hat{\beta}_1$:

$$\begin{aligned}\hat{\beta}_1 &= \frac{(-.1432)(8.2850) - (1.1040)(.0018)}{(8.7018)(8.2850) - (.0000033)} \\ &= -.01649;\end{aligned}$$

now, $\hat{\beta}_2$:

$$\begin{aligned}\hat{\beta}_2 &= \frac{(1.1040)(8.7018) - (.1432)(.0018)}{(8.7018)(8.2850) - (.0000033)} \\ &= .1333;\end{aligned}$$

finally, $\hat{\beta}_0$:

$$\begin{aligned}\hat{\beta}_0 &= .4918 + .01649(.4969) - .1333(.5095) \\ &= .43208.\end{aligned}$$

Taken together, the three parameter estimates gives us the following model (rounding to the third decimal):

$$\hat{Y} = .43 - .016X_1 + .133X_2.$$

Computation is not magical, mystical, or mysterious. It *only* is a function of the means of X_1 , X_2 , and Y , deviations from the mean, and products of deviations from the mean. The exercise is useful, however, for you to see what is going on. The regression coefficients are a function of *all* the pieces of information (all of the data). The coefficients may also be referred to as *partial* regression coefficients because they give the direct "effect" of X_1 on Y , net any effects from X_2 . Note what is happening with X_2 in the estimation of $\hat{\beta}_1$: its influence is being "swept" out or partialled out of the estimate so what is left is the direct effect of X_1 on Y . Similar remarks apply about $\hat{\beta}_2$.

Now that we see where the coefficients come from, let's explore some other issues with these data. First, suppose that some variable, call it X_3 , was some mathematical function of X_1 ? Suppose that $X_3 = 2 \times X_1$; therefore $X_1 = X_3/2$ implying that X_1 is completely determined by X_3 . The correlation in this case would be 1.0.

If we proceeded to estimate a regression model in Stata by typing

```
regress y x1 x3
```

no estimate could be made for $\hat{\beta}_1$ because the denominator in the estimating equation for $\hat{\beta}_1$ would be 0 as the two covariates are perfectly correlated. (Question: what would the slope coefficient be for the following model?):

```
regress x1 x3
```

Answer: .5 (WHY?: Because $X_1 = .5 \times X_3$.) Also, in this model, the r^2 would have to be 1.0. In Stata, you will get estimates for X_3 because Stata simply *drops* X_1 out of the analysis and proceeds as if you were estimating a bivariate model. Indeed, run

```
regress y x3
```

and compare it to the original three-variable model. The coefficient estimates will be identical.

This exercise illustrates the simple point that the estimating equations will not be defined under conditions of perfect correlation; however suppose the correlation is very high, but not perfect.

Let's create a variable called X_4 that is correlated with X_1 at a level of .998 (nearly perfect). Now if we run the model,

```
regress y x1 x4
```

what happens? The estimating equations will be solvable because the denominator for the slope coefficients is not 0 (it would be really close to 0, but not exactly 0). The problem with regression when the covariates are extremely highly correlated is that, as we'll see later, the variance around the parameter estimate drastically increases. The precision of the estimate declines because it is increasingly harder to separate out—or net out—the unique contribution of one variable on Y . Indeed if you estimate two bivariate regressions, notice how the coefficient estimate changes than when compared to the three-variable model.

This is an extreme example of *multicollinearity*. Now before we go crazy worrying about multicollinearity, understand that there is a silver lining here: you *will* get estimates even with correlation between or among the covariates. The quality of the estimates will decrease as the correlation increases, *but* you will get estimates. Later in the semester we will worry more about this issue, and discover that, barring correlations that are exceptionally high, moderately strong correlations between covariates causes little trouble for us. (That's a good thing.)

Now suppose that the correlation between two variables was near 0; that is, the independent variables were close to being independent of one another. In our example, the variable X_2 is correlated with X_1 at a level of .0002—nearly 0. If run the model

```
regress y x1 x2
```

we will get the estimates we computed earlier. However, note from the least squares solutions an interesting implication. If the correlation between two variables is near-0, then the sum of the product of their mean deviations (point out the term) [akin to a covariance] will be 0. Also, the second term in the denominator will tend toward 0 and so the regression coefficient in the multiple regression setting looks similar to the bivariate case. Indeed, if you run

```
regress y x1
```

and

```
regress y x2
```

then you will see that the coefficient estimates from the three-variable model are nearly identical to the coefficient estimates from the two bivariate models. This is the case because the two variables

are nearly independent. This is a good thing in the sense that you can be confident that you're measuring and accounting for two unique relationships. In general, covariates will have a correlation something other than 0, and so the terms in the estimating equations will effect the partial regression coefficient.

Finally, suppose that you were not thinking clearly and you estimated a regression model with a covariate that was a constant. Suppose K was a variable having the same value for all observations. If you ran

```
regress y x1 K
```

you would see that Stata would drop K out of the model because the estimating equation for the parameter estimate would not be estimable. (Why?) However, suppose that instead of being a constant, you had a new variable Q that had limited variance and was close to being a constant? The implication would be that you would get estimates for the parameter estimate; however, because Q had little variance, the error variance around its parameter would be very large because the regression estimate would be working with a limited amount of information regarding Q . The solution: collect data across a wider range of Q .

All of the above holds for the case where you have n dependent variables. The algebra gets a little more cumbersome because you have more terms to deal with; however, it is a straightforward extension. The goodness-of-fit indicators extend in a straightforward way. The standard error of the estimate (which gives the average residual) is given by

$$s.e.(e) = \sqrt{\frac{\sum e_i^2}{n - k - 1}},$$

where k corresponds to the number of slope coefficients estimated and 1 corresponds to the intercept term (these are your lost degrees of freedom). (Note again that this quantity is the RMSE on the Stata output). The r^2 is computed in the same way as before, although with multiple regressors, the square root of the term now represents the simple correlation between the observed Y and the \hat{Y} . (To verify this, run a regression with two or more independent variables; output the predicted values; run a correlation between them; square it and compare it to the model r^2 : it will be identical.)

9 Variance Components

The logic of the least squares solution is to minimize the sum of the squared residuals, that is

$$\sum(e_i)^2.$$

This quantity is one of our variance components and is given by

$$\sum(e_i)^2 = \sum(y_i - \hat{y}_i)^2.$$

Visually, we can see this component by looking at the spread of the data around the regression line. The sum of the squared residuals (or the SSE) is the variance component attributable to “unexplained” variance, that is deviations from the predicted regression function that can't be accounted for by x . Apart from this unexplained variation, we can also think in terms of “explained variance.” Regression analysis provides us with information about how knowledge of x improves our ability to make predictions about y . When x and y are strongly related, then we can use x profitably to make some conclusions about y . Absent x , our best prediction of y would be \bar{y} , simply

because the mean (under suitable conditions) best describes the central tendency of a variable. With x and a regression model, we can ask the question: how much does knowledge of x improve our ability to predict y ?

The answer to this question is given by the second of our two variance components. Visually, we can look at our data and the predicted regression function. Consider the Florida data. (SHOW GRAPH OF DATA).

The deviation for the observed to the predicted points for each county corresponds to error—it is unexplained variation; however, the deviation between the predicted point and the mean of y corresponds to variation accounted for by our model. If we were to guess the mean of y for each county, we would be off in our predictions equal to the amount of distance between the predicted point and the mean. As the gap widens, we can make better guesses by using information about x . The total amount of this variation is given by

$$\Sigma(\hat{y}_i - \bar{y})^2$$

and can be denoted as the sum of the squares due to the regression, or SSR. The two components taken together correspond to the total amount of variance.

Because we can decompose the total variance into two parts, a natural goodness-of-fit indicator emerges. Suppose that

$$\Sigma(y_i - \hat{y}_i)^2 > \Sigma(\hat{y}_i - \bar{y})^2,$$

then the residual variance or error variance would be greater than the explained variance or model variance. As a proportion of the total variance, unexplained variance constitutes a greater share. However, suppose that

$$\Sigma(y_i - \hat{y}_i)^2 < \Sigma(\hat{y}_i - \bar{y})^2,$$

then model variance would be greater than error variance, and thus would comprise a greater proportion of the total variance in the model. To directly measure the proportion of the variance accounted for by the model, we can use a measure known as the r^2 . This is given by

$$1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \hat{y}_i)^2 + \Sigma(\hat{y}_i - \bar{y})^2}$$

or equivalently

$$\frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \hat{y}_i)^2 + \Sigma(\hat{y}_i - \bar{y})^2}$$

or

$$\frac{\text{SSR}}{\text{SSTO}}.$$

As this is a ratio, it is bound between 0 and 1, where 1 denotes the case where variation in x perfectly accounts for variation in y and 0 denotes the absence of a relationship between x and y . (That is, every $\hat{y}_i = \bar{y}$.)

There are good things and bad things about the r^2 indicator. For what it is defined as, it is reasonable to use; however, is it always the case that low r^2 imply a poor model? The answer is “no.” Why? There are instances when you will be interested in the r^2 ; however, I wouldn’t hold it up as the end all to be all.

Another indicator of goodness-of-fit is given by the so-called standard error of the estimate. This is closely related to the r^2 because it is a function of the variance components. To understand what is going on, let’s begin with the residuals. The variance of the residuals is given by

$$s^2(e_i) = \frac{\Sigma(y_i - \hat{y}_i)^2}{n - 2},$$

which may be interpreted, like a standard variance, as the average squared deviation between the observed data and the predicted values. (Note that one may interpret the variance of x as the average of the squared deviations around the mean of x .) Owing to this interpretation, the variance of the residuals is sometimes referred to as the *mean square error*, or the mean of the sum of squares due to error. The denominator is $n - 2$ because in order to derive the variance components, two parameters must be estimated, a and b , thus using up 2 degrees of freedom (think of analogy to the mean and its variance).

The standard error of the regression estimate is nothing more than the standard error of the residuals, which is given by

$$\text{s.e.}(e_i) = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)}{n - 2}},$$

and is interpreted as the average deviation of the observed data around the predicted points (interpreted like a standard deviation). This term is sometimes referred to as the *root mean square error* which makes sense, because it is the square root of the MSE. Another way to think of the s.e. of the estimate is that it is an indicator of the average residual found in the data. As a measure of goodness-of-fit, it is appealing because the magnitude of the deviations are expressed in terms of units of the dependent variable. If we assumed the residuals were normally distributed (approximately), then we could use the 68-95-99.7 percent rule as learned in basic statistics to make some fundamental claims. About 68 percent of the residuals will be in the range plus or minus 354 votes. Clearly there is a connection to the r^2 insofar as the smaller this standard error, the more compact are the observed data to the predicted regression function. Hence, the smaller the standard error of the estimate becomes, the larger the r^2 becomes. Note that the standard error of the estimate can be expressed as

$$\text{s.e.}(e_i) = \sqrt{\frac{SSTO - SSR}{n - 2}}.$$

Since the ratio of SSR to SSTO is the r^2 , it is clear that the standard error of the estimate is a function of the same variance components. Thus, models with low r^2 will have higher standard errors of the estimate. The same caviates apply. (Re-run the Buchanan-Bush model excluding Palm Beach County and see what happens to the variance components).

10 Correlation

Note that there is a close connection to correlation and regression. The simplest connection is in recognizing that when the slope coefficient is positive, the square root of the r^2 is equal to the correlation coefficient; when the slope coefficient is negative, the negative of the square root of the r^2 is equal to the correlation coefficient. Hence, for the Buchanan model, we see that r^2 is .39 and so the square root of this gives r , the correlation coefficient, which is .62.

Note the formula for the correlation coefficient is

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}}$$

which is equivalent to

$$r = \frac{\text{COV}_{xy}}{s_x s_y}.$$

It's important to note close connection between r and b . Recall that we expressed b as

$$b = \frac{\text{COV}_{xy}}{\text{var}_x}.$$

The big difference between r and b is the symmetry in r . Note that we could reverse the term in the denominator and it would make no difference to the outcome. The correlation coefficient gives us the linear associate between two variables. We need not designate one as the dependent variable and one as the independent variable. Hence $r_{xy} = r_{yx}$. However, in the regression model, there is asymmetry in the slope. The denominator is the variance of x . It *would* matter if we substituted in the variance of y . For the slope coefficient, the x and y are not exchangeable.

Nevertheless, the two quantities are related. It's useful to note that the slope coefficient can be written directly in terms of the correlation coefficient:

$$b = r_{xy}s_y/s_x.$$

For the Buchanan model, note that the correlation between Bush and Buchanan is .625. The standard deviation for the Bush vote is 57154.17 and for the Buchanan vote, 449.92. The ratio of the standard deviations is

$$s_{Buchanan}/s_{Bush} = .0079.$$

Multiplying $127.03 \times .625$ gives us .0049, which is identical to the regression output. Note that had we flip-flopped the dependent and independent variables, then we would have a slope coefficient equal to $r_{xy}s_{Bush}/s_{Buchanan}$ which would be $.625 \times 127.03 = 79.39$. Go estimate the regression model using the command

regress Bush Buchanan

and verify that you'll get this as your slope coefficient.