Properties of Regression

"I like cool people, but why should I care?
Because I'm busy tryna fit a line with the least squares"
– J-Wong

Summation operator Now we move on to a review of the least squares estimator.[15] Before we begin, let's introduce some new notation starting with the **summation operator**. The Greek letter $\sum$ (the capital Sigma) denotes the summation operator. Let $x_1, x_2, \ldots, x_n$ be a sequence of numbers. We can compactly write a sum of numbers using the summation operator as:

$$\sum_{i=1}^{n} x_i \equiv x_1 + x_2 + \cdots + x_n$$

The letter i is called the index of summation. Other letters are sometimes used, such as j or k, as indices of summation. The subscripted variable simply represents a specific value of a random variable, x. The numbers 1 and n are the lower limit and upper limit of the summation. The expression $\sum_{i=1}^{n}$ can be stated in words as "sum the numbers x_i for all values of i from 1 to n". An example can help clarify:

$$\sum_{i=6}^{9} x_i = x_6 + x_7 + x_8 + x_9$$

The summation operator has three properties. The first property is called the constant rule. Formally, it is:

$$\text{For any constant } c: \quad \sum_{i=1}^{n} c = nc \tag{13}$$

Let's consider an example. Say that we are given:

$$\sum_{i=1}^{3} 5 = (5 + 5 + 5) = 3 \cdot 5 = 15$$

A second property of the summation operator is:

$$\sum_{i=1}^{n} cx_i = c \sum_{i=1}^{n} x_i \tag{14}$$

[15] This chapter is heavily drawn from Wooldridge [2010] and Wooldridge [2015]. All errors are my own.

Again let's use an example. Say we are given:

$$\begin{aligned}\sum_{i=1}^{3} 5x_i &= 5x_1 + 5x_2 + 5x_3 \\ &= 5(x_1 + x_2 + x_3) \\ &= 5\sum_{i=1}^{3} x_i\end{aligned}$$

We can apply both of these properties to get the following third property:

$$\text{For any constant } a \text{ and } b : \sum_{i=1}^{n}(ax_i + by_i) = a\sum_{i=1}^{n} x_i + b\sum_{j=1}^{n}$$

Before leaving the summation operator, it is useful to also note things which are **not** properties of this operator. Two things which summation operators **cannot** do:

$$\begin{aligned}\sum_{i}^{n} \frac{x_i}{y_i} &\neq \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} y_i} \\ \sum_{i=1}^{n} x_i^2 &\neq \left(\sum_{i=1}^{n} x_i\right)^2\end{aligned}$$

We can use the summation indicator to make a number of calculations, some of which we will do repeatedly over the course of this book. For instance, we can use the summation operator to calculate the **average**:

$$\begin{aligned}\overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \frac{x_1 + x_2 + \cdots + x_n}{n}\end{aligned} \tag{15}$$

where $\overline{x}$ is the average (mean) of the random variable x_i. Another calculation we can make is a random variable's deviations from its own mean. The sum of the deviations from the mean is always equal to zero:

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0 \tag{16}$$

Let's illustrate this with an example in Table 5:

Consider a sequence of two numbers $\{y_1, y_2, \ldots, y_n\}$ and $\{x_1, x_2, \ldots, x_n\}$. Then we may consider double summations over possible values of x's and y's. For example, consider the case where $n = m = 2$. Then,

x	$x - \overline{x}$
10	2
4	-4
13	5
5	-3
Mean=8	Sum=0

Table 5: Sum of deviations equalling zero

$\sum_{i=1}^{2}\sum_{j=1}^{2} x_i y_j$ is equal to $x_1y_1 + x_1y_2 + x_2y_1 + x_2y_2$. This is because:

$$\begin{aligned} x_1y_1 + x_1y_2 + x_2y_1 + x_2y_2 &= x_1(y_1 + y_2) + x_2(y_1 + y_2) \\ &= \sum_{i=1}^{2} x_i(y_1 + y_2) \\ &= \sum_{i=1}^{2} x_i\left(\sum_{j=1}^{2} y_j\right) \\ &= \sum_{i=1}^{2}\left(\sum_{j=1}^{2} x_i y_j\right) \\ &= \sum_{i=1}^{2}\sum_{j=1}^{2} x_i y_j \end{aligned}$$

One result that will be very useful throughout the semester is:

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\overline{x})^2 \tag{17}$$

An overly long, step-by-step, proof is below. Note that the summation index is suppressed after the first line for brevity sake.

$$\begin{aligned} \sum_{i=1}^{n}(x_i - \overline{x})^2 &= \sum_{i=1}^{n}(x_i^2 - 2x_i\overline{x} + \overline{x}^2) \\ &= \sum x_i^2 - 2\overline{x}\sum x_i + n\overline{x}^2 \\ &= \sum x_i^2 - 2\frac{1}{n}\sum x_i \sum x_i + n\overline{x}^2 \\ &= \sum x_i^2 + n\overline{x}^2 - \frac{2}{n}\left(\sum x_i\right)^2 \\ &= \sum x_i^2 + n\left(\frac{1}{n}\sum x_i\right)^2 - 2n\left(\frac{1}{n}\sum x_i\right)^2 \\ &= \sum x_i^2 - n\left(\frac{1}{n}\sum x_i\right)^2 \\ &= \sum x_i^2 - n\overline{x}^2 \end{aligned}$$

A more general version of this result is:

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) &= \sum_{i=1}^{n}x_i(y_i-\overline{y}) \\ &= \sum_{i=1}^{n}(x_i-\overline{x})y_i \\ &= \sum_{i=1}^{n}x_iy_i - n(\overline{xy}) \end{aligned} \tag{18}$$

Or:

$$\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) = \sum_{i=1}^{n}x_i(y_i-\overline{y}) = \sum_{i=1}^{n}(x_i-\overline{x})y_i = \sum_{i=1}^{n}x_iy_i - n(\overline{xy}) \tag{19}$$

Expected value The **expected value** of a random variable, also called the expectation and sometimes the population mean, is simply the weighted average of the possible values that the variable can take, with the weights being given by the probability of each value occurring in the population. Suppose that the variable X can take on values $x_1, x_2, \ldots, x_k$ each with probability $f(x_1), f(x_2), \ldots, f(x_k)$, respectively. Then we define the expected value of X as:

$$\begin{aligned}E(X) &= x_1f(x_1) + x_2f(x_2) + \cdots + x_kf(x_k) \\ &= \sum_{j=1}^{k}x_jf(x_j)\end{aligned} \tag{20}$$

Let's look at a numerical example. If X takes on values of -1, 0 and 2 with probabilities 0.3, 0.3 and 0.4,[16] respectively. Then the expected value of X equals:

[16] Recall the law of total probability requires that all marginal probabilities sum to unity.

$$\begin{aligned}E(X) &= (-1)(0.3) + (0)(0.3) + (2)(0.4) \\ &= 0.5\end{aligned}$$

In fact you could take the expectation of a function of that variable, too, such as X^2. Note that X^2 takes only the values 1, 0 and 4 with probabilities 0.3, 0.3 and 0.4. Calculating the expected value of X^2 therefore is:

$$\begin{aligned}E(X^2) &= (-1)^2(0.3) + (0)^2(0.3) + (2)^2(0.4) \\ &= 1.9\end{aligned}$$

The first property of expected value is that for any constant, c, $E(c) = c$. The second property is that for any two constants, a and b, then $E(aX + b) = E(aX) + E(b) = aE(X) + b$. And the third property is that if we have numerous constants, $a_1, \ldots, a_n$ and many random variables, $X_1, \ldots, X_n$, then the following is true:

$$E(a_1X_1 + \cdots + a_nX_n) = a_1E(X_1) + \cdots + a_nE(X_n)$$

We can also express this using the expectation operator:

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1} a_i E(X_i)$$

And in the special case where $a_i = 1$, then

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

Variance The expectation operator, $E(\cdot)$, is a **population** concept. It refers to the whole group of interest, not just the sample we have available to us. Its intuition is loosely similar to the average of a random variable in the population. Some additional properties for the expectation operator can be explained assuming two random variables, W and H.

$$\begin{aligned} E(aW + b) &= aE(W) + b \text{ for any constants } a, b \\ E(W + H) &= E(W) + E(H) \\ E(W - E(W)) &= 0 \end{aligned}$$

Consider the variance of a random variable, W:

$$V(W) = \sigma^2 = E[(W - E(W))^2] \text{ in the population}$$

We can show that:

$$V(W) = E(W^2) - E(W)^2 \tag{21}$$

In a given sample of data, we can estimate the variance by the following calculation:

$$\widehat{S}^2 = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

where we divide by $n - 1$ because we are making a degrees of freedom adjustment from estimating the mean. But in large samples, this degree of freedom adjustment has no practical effect on the value of $\widehat{S}^2$.[17]

[17] Whenever possible, I try to use the "hat" to represent an estimated statistic. Hence $\widehat{S}^2$ instead of just S^2. But it is probably more common to see the sample variance represented as S^2.

A few more properties of variance. First, the variance of a line is:

$$V(aX + b) = a^2 V(X)$$

And the variance of a constant is zero (i.e., $V(c) = 0$ for any constant, c). The variance of the sum of two random variables is equal to:

$$V(X + Y) = V(X) + V(Y) + 2(E(XY) - E(X)E(Y)) \tag{22}$$

If the two variables are independent, then $E(XY) = E(X)E(Y)$ and $V(X + Y)$ is just equal to the sum of $V(X) + V(Y)$.

Covariance The last part of equation 22 is called the covariance. The **covariance** measures the amount of linear dependence between two random variables. We represent it with the $C(X, Y)$ operator. $C(X, Y) > 0$ indicates that two variables move in the same direction, whereas $C(X, Y) < 0$ indicates they move in opposite directions. Thus we can rewrite Equation 22 as:

$$V(X + Y) = V(X) + V(Y) + 2C(X, Y)$$

While it's tempting to say that a zero covariance means two random variables are unrelated, that is incorrect. They could have a nonlinear relationship. The definition of covariance is

$$C(X, Y) = E(XY) - E(X)E(Y) \tag{23}$$

As we said, if X and Y are independent, then $C(X, Y) = 0$ in the population.[18] The covariance between two linear functions is:

$$C(a_1 + b_1X, a_2 + b_2Y) = b_1b_2C(X, Y)$$

[18] It may be redundant to keep saying this, but since we've been talking about only the population this whole time, I wanted to stress it again for the reader.

The two constants, a_1 and a_2, zero out because their mean is themselves and so the difference equals zero.

Interpreting the magnitude of the covariance can be tricky. For that, we are better served looking at **correlation**. We define correlation as follows. Let $W = \frac{X - E(X)}{\sqrt{V(X)}}$ and $Z = \frac{Y - E(Y)}{\sqrt{V(Y)}}$. Then:

$$Corr(W, Z) = \frac{C(X, Y)}{\sqrt{V(X)V(Y)}} \tag{24}$$

The correlation coefficient is bounded between –1 and 1. A positive (negative) correlation indicates that the variables move in the same (opposite) ways. The closer to 1 or –1 the stronger the linear relationship is.

Population model We begin with cross-sectional analysis. We will also assume that we can collect a random sample from the population of interest. Assume there are two variables, x and y, and we would like to see how y varies with changes in x.[19]

[19] Notice – this is not necessarily causal language. We are speaking first and generally just in terms of two random variables systematically moving together in some measurable way.

There are three questions that immediately come up. One, what if y is affected by factors other than x? How will we handle that? Two, what is the functional form connecting these two variables? Three, if we are interested in the causal effect of x on y, then how can we distinguish that from mere correlation? Let's start with a specific model.

$$y = \beta_0 + \beta_1 x + u \tag{25}$$

This model is assumed to hold in the *population*. Equation 25 defines a **linear bivariate regression model**. For causal inference, the terms

on the left-hand-side are usually thought of as the effect, and the terms on the right-hand-side are thought of as the causes.

Equation 25 explicitly allows for other factors to affect y by including a random variable called the **error** term, u. This equation also explicitly models the functional form by assuming that y is linearly dependent on x. We call the β_0 coefficient the **intercept parameter**, and we call the β_1 coefficient the **slope parameter**. These, note, describe a population, and our goal in empirical work is estimate their values. I will emphasize this several times throughout this book: we never directly observe these parameters, because they are not data. What we can do, though, is estimate these parameters using *data* and *assumptions*. We just have to have credible assumptions to *accurately* estimate these parameters with data. We will return to this point later. In this simple regression framework, all unobserved variables are subsumed by the error term.

First, we make a simplifying assumption without loss of generality. Let the expected value of u be zero in the population. Formally:

$$E(u) = 0 \tag{26}$$

where $E(\cdot)$ is the expected value operator discussed earlier. Normalizing ability to be zero in the population is harmless. Why? Because the presence of β_0 (the intercept term) always allows us this flexibility. If the average of u is different from zero – for instance, say that it's α_0 – then we just adjust the intercept. Adjusting the intercept has no effect on the β_1 slope parameter, though.

$$y = (\beta_0 + \alpha_0) + \beta_1 x + (u - \alpha_0)$$

where $\alpha_0 = E(u)$. The new error term is $u - \alpha_0$ and the new intercept term is $\beta_0 + \alpha_0$. But while those two terms changed, notice what did *not* change: the slope, β_1, has not changed.

Mean independence An assumption that meshes well with our elementary treatment of statistics involves the mean of the error term for each "slice" of the population determined by values of x:

$$E(u|x) = E(u) \text{ for all values } x \tag{27}$$

where $E(u|x)$ means the "expected value of u given x". If equation 27 holds, then we say that u is **mean independent** of x. An example might help here. Let's say we are estimating the effect of schooling on wages, and u is unobserved ability. Mean independence requires that $E[ability|x = 8] = E[ability|x = 12] = E[ability|x = 16]$ so that the average ability is the same in the different portions of the population with an 8^{th} grade education, a 12^{th} grade education and a college

education. Because people choose education, though, based partly on that unobserved ability, equation 27 is almost certainly violated in this actual example.

Combining this new assumption, $E[u|x] = E[u]$ (a non-trivial assumption to make), with $E[u] = 0$ (a normalization and trivial assumption), and you get the following new assumption:

$$E(u|x) = 0, \text{ for all values } x \tag{28}$$

Equation 28 is called the **zero conditional mean assumption** and is a key identifying assumption in regression models. Because the conditional expected value is a linear operator, $E(u|x) = 0$ implies

$$E(y|x) = \beta_0 + \beta_1 x$$

which shows the **population regression function** is a linear function of x, or what Angrist and Pischke [2009] call the conditional expectation function.[20] This relationship is crucial for the intuition of the parameter, β_1, as a *causal parameter*.

[20] Notice that the conditional expectation passed through the linear function leaving a constant, because of the first property of the expectation operator, and a constant times x. This is because the conditional expectation of $E[X|X] = X$. This leaves us with $E[u|X]$ which under zero conditional mean is equal to zero.

Least Squares Given data on x and y, how can we estimate the population parameters, β_0 and β_1? Let $\left\{(x_i, y_i) : i = 1, 2, \ldots, n\right\}$ be a **random** sample of size n (the number of observations) from the population. Plug any observation into the population equation:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

where i indicates a particular observation. We observe y_i and x_i but not u_i. We just know that u_i is there. We then use the two population restrictions that we discussed earlier:

$$\begin{aligned} E(u) &= 0 \\ C(x, u) &= 0 \end{aligned}$$

to obtain estimating equations for β_0 and β_1. We talked about the first condition already. The second one, though, means that x and u are *uncorrelated* because recall covariance is the numerator of correlation equation (equation 24). Both of these conditions imply equation 28:

$$E(u|x) = 0$$

With $E(xu) = 0$, we get $E(u) = 0$, $C(x, u) = 0$. Notice that if $C(x, u) = 0$, it implies x and u are independent.[21] Next we plug in for u, which is equal to $y - \beta_0 - \beta_1 x$:

[21] See equation 23.

$$\begin{aligned} E(y - \beta_0 - \beta_1 x) &= 0 \\ E(x[y - \beta_0 - \beta_1 x]) &= 0 \end{aligned}$$

These are the two conditions in the **population** that effectively determine β_0 and β_1. And again, note that the notation here is population concepts. We don't have access to populations, though we do have their sample counterparts:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i) = 0 \tag{29}$$

$$\frac{1}{n}\sum_{i=1}^{n}(x_i[y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i]) = 0 \tag{30}$$

where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the estimates from the data.[22] These are two linear equations in the two unknowns $\widehat{\beta}_0$ and $\widehat{\beta}_1$. Recall the properties of the summation operator as we work through the following sample properties of these two equations. We begin with equation 29 and pass the summation operator through.

[22] Notice that we are dividing by n, not $n-1$. There is no degrees of freedom correction, in other words, when using samples to calculate means. There is a degrees of freedom correction when we start calculating higher moments.

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i) &= \frac{1}{n}\sum_{i=1}^{n}(y_i) - \frac{1}{n}\sum_{i=1}^{n}\widehat{\beta_0} - \frac{1}{n}\sum_{i=1}^{n}\widehat{\beta_1}x_i \\ &= \frac{1}{n}\sum_{i=1}^{n}y_i - \widehat{\beta_0} - \widehat{\beta_1}\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right) \\ &= \overline{y} - \widehat{\beta_0} - \widehat{\beta_1}\overline{x}\end{aligned}$$

where $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ which is the average of the n numbers $\{y_i : 1, \ldots, n\}$. For emphasis we will call $\overline{y}$ the **sample average**. We have already shown that the first equation equals zero (Equation 29), so this implies $\overline{y} = \widehat{\beta_0} + \widehat{\beta_1}\overline{x}$. So we now use this equation to write the intercept in terms of the slope:

$$\widehat{\beta_0} = \overline{y} - \widehat{\beta_1}\overline{x}$$

We now plug $\widehat{\beta_0}$ into the second equation, $\sum_{i=1}^{n} x_i(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i) = 0$. This gives us the following (with some simple algebraic manipulation):

$$\begin{aligned}\sum_{i=1}^{n}x_i[y_i - (\overline{y} - \widehat{\beta_1}\overline{x}) - \widehat{\beta_1}x_i] &= 0 \\ \sum_{i=1}^{n}x_i(y_i - \overline{y}) &= \widehat{\beta_1}\left[\sum_{i=1}^{n}x_i(x_i - \overline{x})\right]\end{aligned}$$

So the equation to solve is[23]

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \widehat{\beta_1}\left[\sum_{i=1}^{n}(x_i - \overline{x})^2\right]$$

If $\sum_{i=1}^{n}(x_i - \overline{x})^2 \neq 0$, we can write:

$$\begin{aligned}\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ &= \frac{\text{Sample covariance}(x_i, y_i)}{\text{Sample variance}(x_i)}\end{aligned} \tag{31}$$

[23] Recall from much earlier that:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) &= \sum_{i=1}^{n}x_i(y_i - \overline{y}) \\ &= \sum_{i=1}^{n}(x_i - \overline{x})y_i \\ &= \sum_{i=1}^{n}x_iy_i - n(\overline{x}\overline{y})\end{aligned}$$

The previous formula for $\widehat{\beta}_1$ is important because it shows us how to take data that we have and compute the slope estimate. The estimate, $\widehat{\beta}_1$, is commonly referred to as the **ordinary least squares (OLS)** slope estimate. It can be computed whenever the sample variance of x_i isn't zero. In other words, if x_i is not constant across all values of i. The intuition is that the variation in x is what permits us to identify its impact in y. This also means, though, that we cannot determine the slope in a relationship if we observe a sample where everyone has the same years of schooling, or whatever causal variable we are interested in.

Once we have calculated $\widehat{\beta}_1$, we can compute the intercept value, $\widehat{\beta}_0$ as $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$. This is the OLS intercept *estimate* because it is calculated using sample averages. Notice that it is straightforward because $\widehat{\beta}_0$ is linear in $\widehat{\beta}_1$. With computers and statistical programming languages and software, we let our computers do these calculations because even when n is small, these calculations are quite tedious.[24]

[24] Back in the old days, though? Let's be glad that the old days of calculating OLS estimates by hand is long gone.

For any candidate estimates, $\widehat{\beta}_0, \widehat{\beta}_1$, we define a **fitted value** for each i as:

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

Recall that $i = \{1, \ldots, n\}$ so we have n of these equations. This is the value we predict for y_i given that $x = x_i$. But there is prediction error because $y \neq y_i$. We call that mistake the **residual**, and here use the $\widehat{u_i}$ notation for it. So the residual equals:

$$\begin{aligned} \widehat{u_i} &= y_i - \widehat{y_i} \\ \widehat{u_i} &= y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i \end{aligned}$$

Suppose we measure the size of the mistake, for each i, by squaring it. Squaring it will, after all, eliminate all negative values of the mistake so that everything is a positive value. This becomes useful when summing the mistakes if we aren't wanting positive and negative values to cancel one another out. So let's do that: square the mistake and add them all up to get, $\sum_{i=1}^{n} \widehat{u_i}^2$:

$$\begin{aligned} \sum_{i=1}^{n} \widehat{u_i}^2 &= \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 \\ &= \sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i)^2 \end{aligned}$$

This equation is called the **sum of squared residuals** because the residual is $\widehat{u_i} = y_i - \widehat{y}$. But, the residual is based on estimates of the slope and the intercept. We can imagine any number of estimates of those values. But what if our goal is to *minimize* the sum of squared residuals by choosing $\widehat{\beta_0}$ and $\widehat{\beta_1}$? Using calculus, it can be shown that

the solutions to that problem yields parameter estimates that are the same as what we obtained before.

Once we have the numbers $\widehat{\beta}_0$ and $\widehat{\beta}_1$ for a given dataset, we write the **OLS Regression line**:

$$\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1} x \tag{32}$$

Let's consider an example in Stata.

```
  set seed 1
  clear
  set obs 10000
  gen x = rnormal()
  gen u = rnormal()
  gen y = 5.5*x + 12*u
  reg y x
  predict yhat1
  gen yhat2 = 0.0732608 + 5.685033*x
  sum yhat*
  predict uhat1, residual
  gen uhat2=y-yhat2
  sum uhat*
  twoway (lfit y x, lcolor(black) lwidth(medium)) (scatter
y x, mcolor(black) msize(tiny) msymbol(point)), title(OLS
Regression Line)
  rvfplot, yline(0)
```

Run the previous lines verbatim into Stata. Notice that the estimated coefficients – y-intercept and slope parameter – are represented in blue and red below in Figure 3.

Recall that we defined the fitted value as $\widehat{y_i}$ and we defined the residual, $\widehat{u_i}$, as $y_i - \widehat{y_i}$. Notice that the scatter plot relationship between the residuals and the fitted values created a spherical pattern suggesting that they are uncorrelated (Figure 4).

Once we have the estimated coefficients, and we have the OLS regression line, we can predict y (outcome) for any (sensible) value of x. So plug in certain values of x, we can immediately calculate what y will probably be with some error. The value of OLS here lies in how large that error is: OLS minimizes the error for a linear function. In fact, it is the best such guess at y for all linear estimators because it minimizes the prediction error. There's always prediction error, in other words, with any estimator, but OLS is the least worst.

Notice that the intercept is the predicted value of y if and when $x = 0$. Since here that value is 0.0732608, it's a little hard to read, but that's because x and u were random draws and so there's a value of

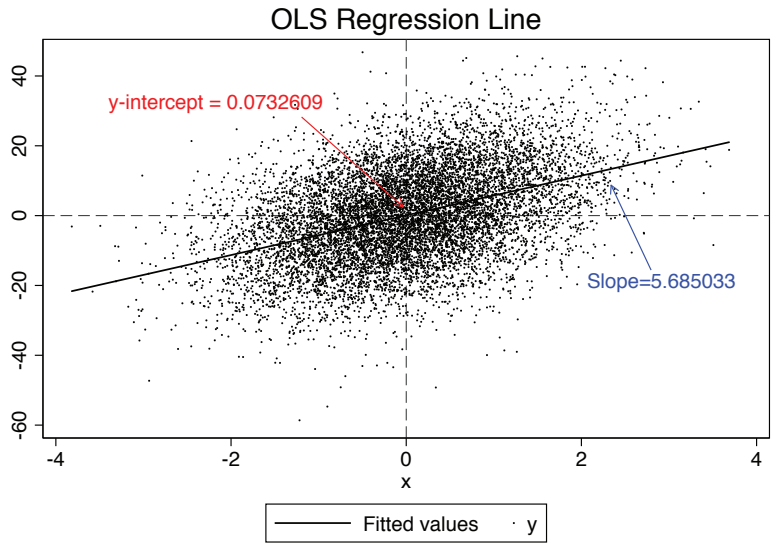

Figure 3: Graphical representation of bivariate regression from y on x

zero for y on average when $x = 0$.[25] The slope allows us to predict changes in y for any reasonable change in x according to:

$$\Delta\widehat{y} = \widehat{\beta}_1 \Delta x$$

And if $\Delta x = 1$, then x increases by one unit, and so $\Delta\widehat{y} = 5.685033$ in our numerical example because $\widehat{\beta}_1 = 5.685033$.

Now that we have calculated $\widehat{\beta}_0$ and $\widehat{\beta}_1$, we get the OLS fitted values by plugging the x_i into the following equation for $i = 1, \ldots, n$:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

The OLS residuals are also calculated by:

$$\widehat{u}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$$

Most residuals will be different from zero (i.e., they do not lie on the regression line). You can see this in Figure 3. Most of the residuals are not on the regression line. Some are positive, and some are negative. A positive residual indicates that the regression line (and hence, the predicted values) underestimates the true value of y_i. And if the residual is negative, then it overestimated.

Algebraic Properties of OLS Remember how we obtained $\widehat{\beta}_0$ and $\widehat{\beta}_1$? When an intercept is included, we have:

$$\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0$$

[25] This is because on average u and x are independent, even if in the sample they aren't. Sample characteristics tend to be slightly different from population properties because of sampling error.

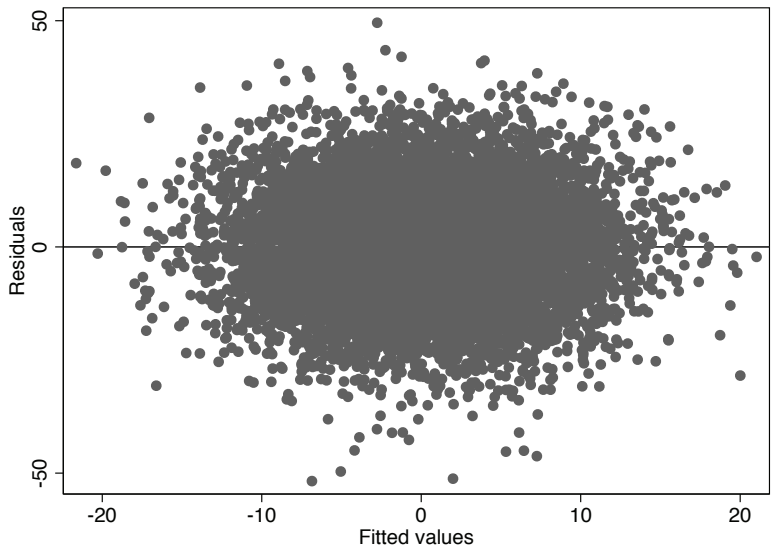

Figure 4: Distribution of residuals around regression line

The OLS residual *always* adds up to zero, by *construction*.

$$\sum_{i=1}^{n} \widehat{u}_i = 0 \tag{33}$$

Sometimes seeing is believing, so let's look at this together. Type the following into Stata verbatim.

```
. clear
. set seed 1234
. set obs 10
. gen x = 9*rnormal()
. gen u = 36*rnormal()
. gen y = 3 + 2*x + u
. reg y x
. predict yhat
. predict residuals, residual
. su residuals
. list
. collapse (sum) x u y yhat residuals
. list
```

Output from this can be summarized in the following table (Table 6).

no.	x	u	y	$\widehat{y}$	$\widehat{u}$	$x\widehat{u}$	$\widehat{y}\widehat{u}$
1.	-4.381653	-32.95803	-38.72134	-3.256034	-35.46531	155.3967	115.4762
2.	-13.28403	-8.028061	-31.59613	-26.30994	-5.28619	70.22192	139.0793
3.	-.0982034	17.80379	20.60738	7.836532	12.77085	-1.254141	100.0792
4.	-.1238423	-9.443188	-6.690872	7.770137	-14.46101	1.790884	-112.364
5.	4.640209	13.18046	25.46088	20.10728	5.353592	24.84179	107.6462
6.	-1.252096	-34.64874	-34.15294	4.848374	-39.00131	48.83337	-189.0929
7.	11.58586	9.118524	35.29023	38.09396	-2.80373	-32.48362	-106.8052
8.	-5.289957	82.23296	74.65305	-5.608207	80.26126	-424.5786	-450.1217
9.	-.2754041	11.60571	14.0549	7.377647	6.677258	-1.838944	49.26245
10.	-19.77159	-14.61257	-51.15575	-43.11034	-8.045414	159.0706	346.8405
Sum	-28.25072	34.25085	7.749418	7.749418	1.91e-06	-6.56e-06	.0000305

Table 6: Simulated data showing the sum of residuals equals zero

Notice the difference between the u, $\widehat{y}$ and $\widehat{u}$ columns. When we sum these ten lines, neither the error term nor the fitted values of y sum to zero. But the residuals *do sum to zero*. This is, as we said, one of the algebraic properties of OLS – coefficients were optimally chosen to ensure that the residuals sum to zero.

Because $y_i = \widehat{y}_i + \widehat{u}_i$ by definition (which we can also see in the above table), we can take the sample average of both sides

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} \widehat{y}_i + \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i$$

and so $\overline{y} = \overline{\widehat{y}}$ because the residuals sum to zero. Similarly, the way that we obtained our estimates yields,

$$\sum_{i=1}^{n} x_i(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0$$

The sample covariance (and therefore the sample correlation) between the explanatory variables and the residuals is always zero (see Table 6):

$$\sum_{i=1}^{n} x_i \widehat{u}_i = 0$$

Because the $\widehat{y}_i$ are linear functions of the x_i, the fitted values and residuals are uncorrelated too (See Table 6):

$$\sum_{i=1}^{n} \widehat{y}_i \widehat{u}_i = 0$$

Both properties hold by construction. In other words, $\widehat{\beta}_0$ and $\widehat{\beta}_1$ were selected *to make them true*.[26]

A third property is that if we plug in the average for x, we predict the sample average for y. That is, the point, $(\overline{x}, \overline{y})$ is on the OLS regression line, or:

$$\overline{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{x}$$

[26] Using the Stata code from Table 6, you can show all these algebraic properties yourself. I encourage you to do so by creating new variables equalling the product of these terms and collapsing as we did with the other variables. This will help you believe these algebraic properties hold.

Goodness of Fit For each observation, we write

$$y_i = \widehat{y}_i + \widehat{u}_i$$

Define the **total** (SST), **explained** (SSE) and **residual** (SSR) sum of squares as

$$SST = \sum_{i=1}^{n}(y_i - \overline{y})^2 \tag{34}$$

$$SSE = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 \tag{35}$$

$$SSR = \sum_{i=1}^{n}\widehat{u}_i^2 \tag{36}$$

These are sample variances when divided by $n-1$.[27] $\frac{SST}{n-1}$ is the sample variance of y_i, $\frac{SSE}{n-1}$ is the sample variance of $\widehat{y}_i$, and $\frac{SSR}{n-1}$ is the sample variance of $\widehat{u}_i$. With some simple manipulation rewrite equation 34:

[27] Recall the earlier discussion about degrees of freedom correction.

$$\begin{aligned} SST &= \sum_{i=1}^{n}(y_i - \overline{y})^2 \\ &= \sum_{i=1}^{n}\left[(y_i - \widehat{y}_i) - (\widehat{y}_i - \overline{y})\right]^2 \\ &= \sum_{i=1}^{n}\left[\widehat{u}_i - (\widehat{y}_i - \overline{y})\right]^2 \end{aligned}$$

And then using that the fitted values are uncorrelated with the residuals (equation 34), we can show that:

$$SST = SSE + SSR$$

Assuming $SST > 0$, we can define the fraction of the total variation in y_i that is explained by x_i (or the OLS regression line) as

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

which is called the **R-squared** of the regression. It can be shown to be equal to the *square* of the correlation between y_i and $\widehat{y}_i$. Therefore $0 \leq R^2 \leq 1$. An R-squared of zero means no linear relationship between y_i and x_i and an R-squared of one means a perfect linear relationship (e.g., $y_i = x_i + 2$). As R^2 increases, the y_i are closer and closer to falling on the OLS regression line.

You don't want to fixate on R^2 in causal inference, though. It's a useful summary measure but it does not tell us about causality. Remember, we aren't trying to explain y; we are trying to estimate causal effects. The R^2 tells us how much of the variation in y_i is explained by the explanatory variables. But if we are interested in the causal effect of a single variable, R^2 is irrelevant. For causal inference, we need equation 28.

Expected Value of OLS Up to now, we motivated simple regression using a population model. But our analysis has been purely algebraic based on a sample of data. So residuals always average to zero when we apply OLS to a sample, regardless of any underlying model. But now our job gets tougher. Now we have to study the statistical properties of the OLS estimator, referring to a population model and assuming random sampling.[28]

[28] This section is a review of a traditional econometrics pedagogy. We cover it for the sake of completeness, as traditionally, econometricians motivated their discuss of causality through ideas like unbiasedness and consistency.

Mathematical statistics is concerned with questions like "how do our estimators behave across different samples of data?" On average, for instance, will we get the right answer if we could repeatedly sample? We need to find the expected value of the OLS estimators – in effect the average outcome across all possible random samples – and determine if we are right on average. This leads naturally to a characteristic called **unbiasedness**, which is a desirable characteristic of all estimators.

$$E(\widehat{\beta}) = \beta \tag{37}$$

Remember our objective is to estimate β_1, which is the slope **population** parameter that describes the relationship between y and x. Our estimate, $\widehat{\beta_1}$ is an **estimator** of that parameter obtained for a specific sample. Different samples will generate different estimates ($\widehat{\beta_1}$) for the "true" (and unobserved) β_1. Unbiasedness is the idea that if we could take as many random samples on Y as we want from the population and compute an estimate each time, the average of these estimates would be equal to β_1.

There are several assumptions required for OLS to be unbiased. We will review those now. The first assumption is called "linear in the parameters". Assume a population model of:

$$y = \beta_0 + \beta_1 x + u$$

where β_0 and β_1 are the unknown population parameters. We view x and u as outcomes of random variables generated by some data generating process. Thus, since y is a function of x and u, both of which are random, then y is also random. Stating this assumption formally shows our goal is to estimate β_0 and β_1.

Our second assumption is "random sampling". We have a random sample of size n, $\{(x_i, y_i) : i = 1, \ldots, n\}$, following the population model. We know how to use this data to estimate β_0 and β_1 by OLS. Because each i is a draw from the population, we can write, for each i:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Notice that u_i here is the unobserved error for observation i. It is *not* the residual that we compute from the data.

The third assumption is called the "sample variation in the explanatory variable". That is, the sample outcomes on x_i are not all the same value. This is the same as saying the sample variance of x is not zero. In practice, this is no assumption at all. If the x_i are all the same value (i.e., constant), we cannot learn how x affects y in the population. Recall that OLS is the covariance of y and x divided by the variance in x and so if x is constant, then we are dividing by zero, and the OLS estimator is undefined.

The fourth assumption is where our assumptions start to have real teeth. It is called the "zero conditional mean" assumption and is probably the most critical assumption in causal inference. In the population, the error term has zero mean given any value of the explanatory variable:

$$E(u|x) = E(u) = 0$$

This is the key assumption for showing that OLS is unbiased, with the zero value being of no importance once we assume $E(u|x)$ does not change with x. Note that we can compute the OLS estimates whether or not this assumption holds, or even if there is an underlying population model.[29]

So, how do we show $\widehat{\beta_1}$ is an unbiased estimate of β_1 (Equation 37)? We need to show that under the four assumptions we just outlined, the expected value of $\widehat{\beta_1}$, when averaged across random samples, will center on β_1. In other words, unbiasedness has to be understood as related to repeated sampling. We will discuss the answer as a series of steps.

Step 1: Write down a formula for $\widehat{\beta_1}$. It is convenient to use the $\frac{C(x,y)}{V(x)}$ form:

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Now get rid of some of this notational clutter by defining $\sum_{i=1}^{n}(x_i - \overline{x})^2 = SST_x$ (i.e., the total variation in the x_i). Rewrite as:

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{SST_x}$$

Step 2: Replace each y_i with $y_i = \beta_0 + \beta_1 x_i + u_i$ which uses the first linear assumption and the fact that we have sampled data (our

[29] We will focus on $\widehat{\beta_1}$. There are a few approaches to showing unbiasedness. One explicitly computes the expected value of $\widehat{\beta_1}$ conditional on x, $\{x_i : i = 1, \ldots, n\}$. Even though this is the more proper way to understand the problem, technically we can obtain the same results by treating the conditioning variables as if they were fixed in repeated samples. That is, to treat the x_i as nonrandom in the derivation. So, the randomness in $\widehat{\beta_1}$ comes through the u_i (equivalently, the y_i). Nevertheless, it is important to remember that x are random variables and that we are taking expectations conditional on knowing them. The approach that we're taking is called sometimes "fixed in repeated samples", and while not realistic in most cases, it gets us to the same place. We use it as a simplifying device because ultimately this chapter is just meant to help you understand this traditional pedagogy better.

second assumption). The numerator becomes:

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\overline{x})y_i &= \sum_{i=1}^{n}(x_i-\overline{x})(\beta_0+\beta_1 x_i+u_i)\\ &= \beta_0\sum_{i=1}^{n}(x_i-\overline{x})+\beta_1\sum_{i=1}^{n}(x_i-\overline{x})x_i+\sum_{i=1}^{n}(x_i+\overline{x})u_i\\ &= 0+\beta_1\sum_{i=1}^{n}(x_i-\overline{x})^2+\sum_{i=1}^{n}(x_i-\overline{x})u_i\\ &= \beta_1 SST_x+\sum_{i=1}^{n}(x_i-\overline{x})u_i\end{aligned}$$

Note, we used $\sum_{i=1}^{n}(x_i-\overline{x})=0$ and $\sum_{i=1}^{n}(x_i-\overline{x})x_i=\sum_{i=1}^{n}(x_i-\overline{x})^2$ to do this.[30]

[30] Told you we would use this result a lot.

We have shown that:

$$\begin{aligned}\widehat{\beta_1} &= \frac{\beta_1 SST_x+\sum_{i=1}^{n}(x_i-\overline{x})u_i}{SST_x}\\ &= \beta_1+\frac{\sum_{i=1}^{n}(x_i-\overline{x})u_i}{SST_x}\end{aligned}$$

Note how the last piece is the slope coefficient from the OLS regression of u_i on $x_i, i:1,\ldots,n$.[31] We cannot do this regression because the u_i are not observed. Now define $w_i=\frac{(x_i-\overline{x})}{SST_x}$ so that we have the following:

[31] I find it interesting that we see so many $\frac{cov}{var}$ terms when working with regression. It shows up constantly. Keep your eyes peeled.

$$\widehat{\beta_1}=\beta_1+\sum_{i=1}^{n}w_i u_i$$

Note the following things that this showed: first, $\widehat{\beta_1}$ is a linear function of the unobserved errors, u_i. The w_i are all functions of $\{x_1,\ldots,x_n\}$. Second, the random difference between β_1 and the estimate of it, $\widehat{\beta_1}$, is due to this linear function of the unobservables.

Step 3: Find $E(\widehat{\beta_1})$. Under the random sampling assumption and the zero conditional mean assumption, $E(u_i|x_1,\ldots,x_n)=0$, that means conditional on each of the x variables:

$$E(w_i u_i|x_1,\ldots,x_n)=w_i E(u_i|x_1,\ldots,x_n)=0$$

because w_i is a function of $\{x_1,\ldots,x_n\}$. This would be true if in the population u and x are correlated.

Now we can complete the proof: conditional on $\{x_1,\ldots,x_n\}$,

$$\begin{aligned} E(\widehat{\beta_1}) &= E\left(\beta_1 + \sum_{i=1}^{n} w_i u_i\right) \\ &= \beta_1 + \sum_{i=1}^{n} E(w_i u_i) \\ &= \beta_1 + \sum_{i=1}^{n} w_i E(u_i) \\ &= \beta_1 + 0 \\ &= \beta_1 \end{aligned}$$

Remember, β_1 is the fixed constant in the population. The estimator, $\widehat{\beta_1}$, varies across samples and is the random outcome: before we collect our data, we do not know what $\widehat{\beta_1}$ will be. Under the four aforementioned assumptions, $E(\widehat{\beta_0}) = \beta_0$ and $E(\widehat{\beta_1}) = \beta_1$.

I find it helpful to be concrete when we work through exercises like this. So let's visualize this in Stata. Let's create a Monte Carlo simulation in Stata. We have the following population model:

$$y = 3 + 2x + u \tag{38}$$

where $x \sim Normal(0, 9)$, $u \sim Normal(0, 36)$. Also, x and u are independent. The following Monte Carlo simulation will estimate OLS on a sample of data 1,000 times. The true β parameter equals 2. But what will the average $\widehat{\beta}$ equal when we use repeated sampling?

```
clear all
program define ols, rclass
version 14.2
syntax [, obs(integer 1) mu(real 0) sigma(real 1) ]

 clear
 drop _all
 set obs 10000
 gen x = 9*rnormal()
 gen u = 36*rnormal()
 gen y = 3 + 2*x + u
 reg y x
 end

simulate beta=_b[x], reps(1000):  ols
su
hist beta
```

Table 7 gives us the mean value of $\widehat{\beta_1}$ over the 1,000 repetitions (repeated sampling). While each sample had a different estimate, the

Variable	Obs	Mean	St. Dev.
beta	1,000	2.000737	0.0409954

Table 7: Monte Carlo simulation of OLS

average for $\widehat{\beta_1}$ was 2.000737, which is close to the true value of 2 (see Equation 38). The standard deviation in this estimator was 0.0409954, which is close to the standard error recorded in the regression itself.[32] Thus we see that the estimate is the mean value of the coefficient from repeated sampling, and the standard error is the standard deviation from that repeated estimation. We can see the distribution of these coefficient estimates in Figure 5.

[32] The standard error I found from running on one sample of data was 0.0393758.

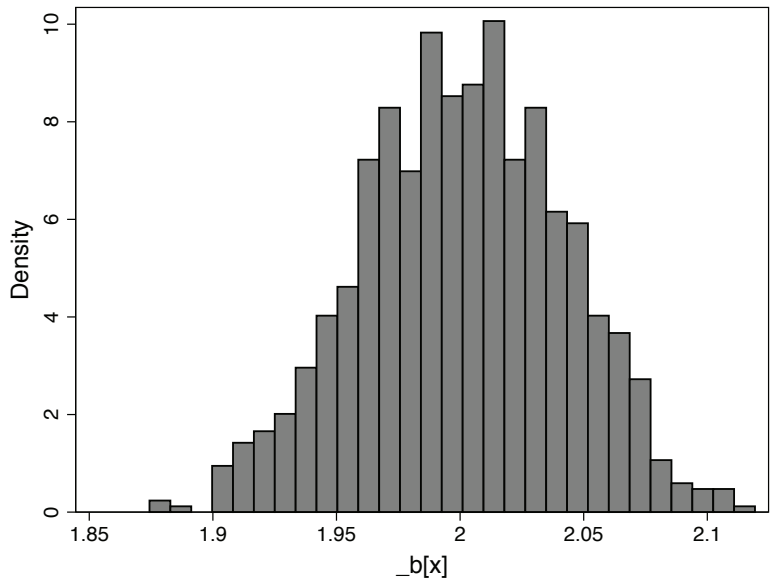

Figure 5: Distribution of coefficients from Monte Carlo simulation.

The problem is, we don't know which kind of sample we have. Do we have one of the "almost exactly 2" samples or do we have one of the "pretty different from 2" samples? We can never know whether we are close to the population value. We hope that our sample is "typical" and produces a slope estimate close to $\widehat{\beta_1}$ but we can't know. Unbiasedness is a property of the procedure of the rule. It is not a property of the estimate itself. For example, say we estimated an 8.2% return on schooling. It is tempting to say 8.2% is an "unbiased estimate" of the return to schooling, but that's incorrect technically. The rule used to get $\widehat{\beta_1} = 0.082$ is unbiased (if we believe that u is unrelated to schooling) – not the actual estimate itself.

Law of iterated expectations As we said earlier in this chapter, the conditional expectation function (CEF) is the mean of some outcome y with some covariate x held fixed. Now we focus more intently on this function.[33] Let's get the notation and some of the syntax out of the way. As noted earlier, we write the CEF as $E(y_i|x_i)$. Note that the CEF is explicitly a function of x_i. And because x_i is random, the CEF is random – although sometimes we work with particular values for x_i, like $E(y_i|x_i = 8 \text{ years schooling})$ or $E(y_i|x_i = \text{Female})$. When there are treatment variables, then the CEF takes on two values: $E(y_i|d_i = 0)$ and $E(y_i|d_i = 1)$. But these are special cases only.

[33] This section is based heavily on Angrist and Pischke [2009].

An important complement to the CEF is the law of iterated expectations (LIE). This law says that an unconditional expectation can be written as the unconditional average of the CEF. In other words $E(y_i) = E\{E(y_i|x_i)\}$. This is a fairly simple idea to grasp. What it states is that if you want to know the unconditional expectation of some random variable y, you can simply calculate the weighted sum of all conditional expectations with respect to some covariate x. Let's look at an example. Let's say that average GPA for females is 3.5, average GPA for males is a 3.2, half the population is females, and half is males. Then:

$$\begin{aligned} E[GPA] &= E\{E(GPA_i|Gender_i)\} \\ &= (0.5 \times 3.5) + (3.2 \times 0.5) \\ &= 3.35 \end{aligned}$$

You probably use LIE all the time and didn't even know it. The proof is not complicated. Let x_i and y_i each be continuously distributed. The joint density is defined as $f_{xy}(u,t)$. The conditional distribution of y given $x = u$ is defined as $f_y(t|x_i = u)$. The marginal densities are $g_y(t)$ and $g_x(u)$.

$$\begin{aligned} E\{E(y|x)\} &= \int E(y|x = u)g_x(u)du \\ &= \int \left[\int t f_{y|x}(t|x = u)dt \right] g_x(u)du \\ &= \int \int t f_{y|x}(t|x = u)g_x(u)dudt \\ &= \int t \left[\int f_{y|x}(t|x = u)g_x(u)du \right] dt \\ &= \int t[f_{x,y}du]dt \\ &= \int t g_y(t)dt \\ &= E(y) \end{aligned}$$

The first line uses the definition of expectation. The second line uses

the definition of conditional expectation. The third line switches the integration order. The fourth line uses the definition of joint density. The sixth line integrates joint density over the support of x which is equal to the marginal density of y. So restating the law of iterated expectations: $E(y_i) = E\{E(y|x_i)\}$.

CEF Decomposition Property The first property of the CEF we will discuss is the CEF Decomposition Property. The power of LIE comes from the way it breaks a random variable into two pieces – the CEF and a residual with special properties. The CEF Decomposition Property states that

$$y_i = E(y_i|x_i) + \varepsilon_i$$

where (i) ε_i is mean independent of x_i, that is

$$E(\varepsilon_i|x_i) = 0$$

and (ii) ε_i is uncorrelated with any function of x_i.

The theorem says that any random variable y_i can be decomposed into a piece that is "explained by x_i" (the CEF) and a piece that is left over and orthogonal to any function of x_i. The proof is provided now. I'll prove the (i) part first. Recall that $\varepsilon_i = y_i - E(y_i|x_i)$ as we will make a substitution in the second line below.

$$\begin{aligned} E(\varepsilon_i|x_i) &= E(y_i - E(y_i|x_i)|x_i) \\ &= E(y_i|x_i) - E(y_i|x_i) \\ &= 0 \end{aligned}$$

The second part of the theorem states that ε_i is uncorrelated with any function of x_i. Let $h(x_i)$ be any function of x_i. Then $E(h(x_i)\varepsilon_i) = E\{h(x_i)E(\varepsilon_i|x_i)\}$ The second term in the interior product is equal to zero by mean independence.[34]

[34] Let's take a concrete example of this proof. Let $h(x_i) = \alpha + \gamma x_i$. Then take the joint expectation $E(h(x_i)\varepsilon_i) = E([\alpha + \gamma x_i]\varepsilon_i)$ Then take conditional expectations $E(\alpha|x_i) + E(\gamma|x_i)E(x_i|x_i)E(\varepsilon|x_i)\} = \alpha + x_i E(\varepsilon_i|x_i) = 0$ after we pass the conditional expectation through.

CEF Prediction Property The second property is the CEF Prediction Property. This states that $E(y_i|x_i) = \arg\min_{m(x_i)} E[(y - m(x_i))^2]$where $m(x_i)$ is any function of x_i. In words, this states that the CEF is the minimum mean squared error of y_i given x_i. By adding $E(y_i|x_i) - E(y_i|x_i) = 0$ to the right hand side we get

$$[y_i - m(x_i)]^2 = [(y_i - E[y_i|x_i]) + (E(y_i|x_i) - m(x_i))]^2$$

I personally find this easier to follow with simpler notation. So replace this expression with the following terms:

$$(a - b + b - c)^2$$

Distribute the terms, rearrange, and replace the terms with their original values until you get the following

$$\arg\min (y_i - E(y_i|x_i))^2 + 2(E(y_i|x_i) - m(x_i)) \times (y_i - E(y_i|x_i)) + (E(y_i|x_i) + m(x_i))^2$$

Now minimize the function with respect to $m(x_i)$. When minimizing this function with respect to $m(x_i)$, note that the first term $(y_i - E(y_i|x_i))^2$ doesn't matter because it does not depend on $m(x_i)$. So it will zero out. The second and third terms, though, do depend on $m(x_i)$. So rewrite $2(E(y_i|x_i) - m(x_i))$ as $h(x_i)$. Also set ε_i equal to $[y_i - E(y_i|x_i)]$ and substitute

$$\arg\min \varepsilon_i^2 + h(x_i)\varepsilon_i + [E(y_i|x_i) + m(x_i)]^2$$

Now minimizing this function and setting it equal to zero we get

$$h'(x_i)\varepsilon_i$$

which equals zero by the Decomposition Property.

ANOVA Theory The final property of the CEF that we will discuss is the analysis of variance theorem, or ANOVA. It is simply that the unconditional variance in some random variable is equal to the variance in the conditional expectation plus the expectation of the conditional variance, or

$$V(y_i) = V[E(y_i|x_i)] + E[V(y_i|x_i)]$$

where V is the variance and $V(y_i|x_i)$ is the conditional variance.

Linear CEF Theorem Angrist and Pischke [2009] give several arguments as to why linear regression may be of interest to a practitioner even if the underlying CEF itself is not linear. I will review of those linear theorems now. These are merely arguments to justify the use of linear regression models to approximate the CEF.[35]

The Linear CEF Theorem is the most obvious theorem of the three that Angrist and Pischke [2009] discuss. Suppose that the CEF itself is linear. Then the population regression is equal to the CEF. This simply states that you should use the population regression to estimate the CEF when you know that the CEF is linear. The proof is provided. If $E(y_i|x_i)$ is linear, then $E(y_i|x_i) = x'\widehat{\beta}$ for some K vector $\widehat{\beta}$. By the Decomposition Property

$$E(x(y - E(y|x)) = E(x(y - x'\widehat{\beta})) = 0$$

Solve this and get $\widehat{\beta} = \beta$. Hence $E(y|x) = x'\beta$.

[35] Note, Angrist and Pischke [2009] make their arguments for using regression, not based on unbiasedness and the four assumptions that we discussed, but rather because regression approximates the CEF. I want to emphasize that this is a subtly different direction. I included the discussion of unbiasedness, though, to be exhaustive. Just note, there is a slight change in pedagogy though.

Best Linear Predictor Theorem Recall that the CEF is the minimum mean squared error predictor of y given x in the class of all functions according to the CEF prediction property. Given this, the population regression function, $E(X'Y)E(X'X)^{-1}$ is the best that we can do in the class of all linear functions.[36] Proof: β solves the population minimum mean squared error problem.

[36] Note that $E(X'Y)E(X'X)^{-1}$ is the matrix notation expression of the population regression, or what we have discussed as $\frac{C(X,Y)}{V(X)}$.

Regression CEF Theorem The function $X\beta$ provides the minimum mean squared error linear approximation to the CEF. That is,

$$\beta = \arg\min_b E\{[E(y_i|x_i) - x_i'b]^2\}$$

Regression anatomy theorem In addition to our discussion of the CEF and regression theorems, we now dissect the regression itself. Here we discuss the **regression anatomy theorem**. The regression anatomy theorem is based on earlier work by Frisch and Waugh [1933] and Lovell [1963].[37] I find it more intuitive when thinking through a specific example and offering up some data visualization. In my opinion, the theorem helps us interpret the individual coefficients of a multiple linear regression model. Say that we are interested in the causal effect of family size on labor supply. We want to regress labor supply onto family size:

[37] A helpful proof of the Frisch-Waugh-Lovell theorem can be found at Lovell [2008].

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

where Y is labor supply and X is family size.

If family size is truly random, then the number of kids is uncorrelated with the unobserved error term. This implies that when we regress labor supply onto family size, our estimate $\widehat{\beta}_1$ can be interpreted as the causal effect of family size on labor supply. Visually, we could just plot the regression coefficient in a scatter plot showing all i pairs of data, and the slope coefficient would be the best fit of this data through this data cloud. That slope would tell us the average causal effect of family size on labor supply.

But how do we interpret $\widehat{\beta}_1$ if the family size is *not random*? After all, we know from living on planet Earth and having even half a brain that a person's family size is usually chosen, not randomly assigned to them. And oftentimes, it's chosen according to something akin to an optimal stopping rule. People pick both the number of kids to have, as well as when to have them, and in some instance, even attempt to pick the gender, and this is all based on a variety of observed and unobserved economic factors that are directly correlated with the decision to supply labor. In other words, using the language we've been using up til now, it's unlikely that $E(u|X) = E(u) = 0$.

But let's say that we have reason to think that the number of kids is *conditionally* random. That is, for a given person of a certain race and age, any remaining variation in family size across a population is random.[38] Then we have the following population model:

[38] Almost certainly not a credible assumption, but stick with me.

$$Y_i = \beta_0 + \beta_1 X_i + \gamma_1 R_i + \gamma_2 A_i + u_i$$

where Y is labor supply, X is family size, R is race, A is age, and u is the population error term.

If we want to estimate the average causal effect of family size on labor supply, then we need two things. First, we need a sample of *data* containing all four of these variables. Without all four of the variables, we cannot estimate this regression model. And secondly, we need for number of kids, X, to be randomly assigned for a given set of race/age.

Now how do we interpret $\widehat{\beta}_1$? And for those who like pictures, how might we visualize this coefficient given there's six dimensions to the data? The regression anatomy theorem tells us both what this coefficient estimate actually means, and it also lets us visualize the data in only two dimensions.

To explain the intuition of the regression anatomy theorem, let's write down a population model with multiple variables. Assume that your main multiple regression model of interest is

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i \tag{39}$$

Now assume an *auxiliary* regression in which the variable x_{1i} is regressed on all the remaining independent variables

$$x_{1i} = \gamma_0 + \gamma_{k-1} x_{k-1i} + \gamma_{k+1} x_{k+1i} + \cdots + \gamma_K x_{Ki} + f_i \tag{40}$$

and $\tilde{x}_{1i} = x_{1i} - \widehat{x}_{1i}$ being the residual from that auxiliary regression. Then the parameter β_1 can be rewritten as:

$$\beta_1 = \frac{C(y_i, \tilde{x}_i)}{V(\tilde{x}_i)} \tag{41}$$

Notice that again we see the coefficient estimate being a scaled covariance, only here the covariance is with respect to the outcome and residual from the auxiliary regression and the scale is the variance of that same residual.

To prove the theorem, note that $E[\tilde{x}_{ki}] = E[x_{ki}] - E[\widehat{x}_{ki}] = E[f_i]$, and plug y_i and residual $\tilde{x}_{ki}$ from x_{ki} auxiliary regression into the covariance $cov(y_i, x_{ki})$.

$$\begin{aligned} \beta_k &= \frac{cov(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i, \tilde{x}_{ki})}{var(\tilde{x}_{ki})} \\ &= \frac{cov(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i, f_i)}{var(f_i)} \end{aligned}$$

Since by construction $E[f_i] = 0$, it follows that the term $\beta_0 E[f_i] = 0$. Since f_i is a linear combination of all the independent variables with the exception of x_{ki}, it must be that

$$\beta_1 E[f_i x_{1i}] = \cdots = \beta_{k-1} E[f_i x_{k-1i}] = \beta_{k+1} E[f_i x_{k+1i}] = \cdots = \beta_K E[f_i x_{Ki}] = 0$$

Consider now the term $E[e_i f_i]$. This can be written as

$$\begin{aligned} E[e_i f_i] &= E[e_i f_i] \\ &= E[e_i \tilde{x}_{ki}] \\ &= E[e_i(x_{ki} - \widehat{x}_{ki})] \\ &= E[e_i x_{ki}] - E[e_i \tilde{x}_{ki}] \end{aligned}$$

Since e_i is uncorrelated with any independent variable, it is also uncorrelated with x_{ki}. Accordingly, we have $E[e_i x_{ki}] = 0$. With regard to the second term of the subtraction, substituting the predicted value from the x_{ki} auxiliary regression, we get

$$E[e_i \tilde{x}_{ki}] = E[e_i(\widehat{\gamma}_0 + \widehat{\gamma}_1 x_{1i} + \cdots + \widehat{\gamma}_{k-1i} + \widehat{\gamma}_{k+1} x_{k+1i} + \cdots + \widehat{x}_K x_{Ki})]$$

Once again, since e_i is uncorrelated with any independent variable, the expected value of the terms is equal to zero. Then it follows that $E[e_i f_i] = 0$.

The only remaining term then is $[\beta_k x_{ki} f_i]$ which equals $E[\beta_k x_{ki} \tilde{x}_{ki}]$ since $f_i = \tilde{x}_{ki}$. The term x_{ki} can be substituted using a rewriting of the auxiliary regression model, x_{ki}, such that

$$x_{ki} = E[x_{ki}|X_{-k}] + \tilde{x}_{ki}$$

This gives

$$\begin{aligned} E[\beta_k x_{ki} \tilde{x}_{ki}] &= \beta_k E[\tilde{x}_{ki}(E[x_{ki}|X_{-k}] + \tilde{x}_{ki})] \\ &= \beta_k \{E[\tilde{x}_{ki}^2] + E[(E[x_{ki}|x_{-k}]\tilde{x}_{ki})]\} \\ &= \beta_k var(\tilde{x}_{ki}) \end{aligned}$$

which follows directly from the orthogonality between $E[x_{ki}|X_{-k}]$ and $\tilde{x}_{ki}$. From previous derivations we finally get

$$cov(y_i, \tilde{x}_{ki}) = \beta_k var(\tilde{x}_{ki})$$

which completes the proof.

I find it helpful to visualize things. Let's look at an example in Stata.

```
. ssc install reganat, replace
. sysuse auto.dta, replace
. regress price length
. regress price length weight headroom mpg
. reganat price length weight headroom mpg, dis(length) biline
```

Let's walk through the regression output. The first regression of

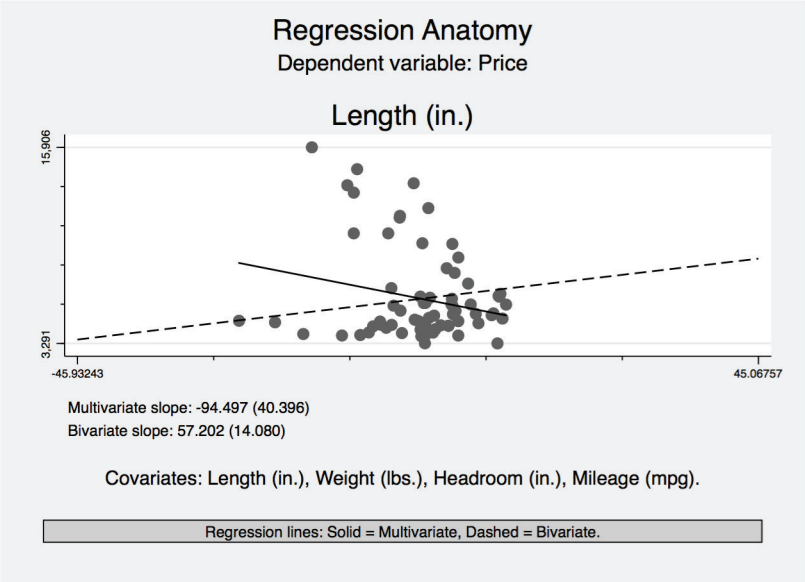

Figure 6: Regression anatomy display.

price on length yields a coefficient of 57.20 on length. But notice the output from the fourth line. The effect on length is -94.5. The first regression is a bivariate regression and gives a positive slope, but the second regression is a multivariate regression and yields a negative slope.

One of the things we can do with regression anatomy (though this isn't its main purpose) is visualize this negative slope from the multivariate regression in nevertheless two dimensional space. Now how do we visualize this first multivariate slope coefficient, given our data has four dimensions? We run the auxiliary regression, use the residuals, and then calculate the slope coefficient as $\frac{cov(y_i, \tilde{x}_i)}{var(\tilde{x}_i)}$. We can also show scatter plots of these auxiliary residuals paired with their outcome observations and slice the slope through them (Figure 6). Notice that this is a useful way to preview the multidimensional correlation between two variables from a multivariate regression.

And as we discussed before, the solid black line is negative while the slope from the bivariate regression is positive. The regression anatomy theorem shows that these two estimators – one being a multivariate OLS and the other being a bivariate regression price and a residual – are identical.

Variance of the OLS Estimators In this chapter we discuss inference under a variety of situations. Under the four assumptions we mentioned earlier, the OLS estimators are unbiased. But these assumptions are not sufficient to tell us anything about the variance in the estimator itself. These assumptions help inform our beliefs that the estimated coefficients, on average, equal the parameter values themselves. But to speak intelligently about the variance of the estimator, we need a measure of dispersion, or spread, in the sampling distribution of the estimators. As we've been saying, this leads us to the variance and ultimately the standard deviation. We could characterize the variance of the OLS estimators under the four assumptions. But for now, it's easiest to introduce an assumption that simplifies the calculations. We'll keep the assumption ordering we've been using and call this the fifth assumption.

The fifth assumption is the *homoskedasticity* or constant variance assumption. This assumption stipulates that our population error term, u, has the same variance given any value of the explanatory variable, x. Formally, it's:

$$V(u|x) = \sigma^2 > 0 \tag{42}$$

where σ is some finite, positive number. Because we assume the zero conditional mean assumption, whenever we assume homoskedasticity, we can also write:

$$E(u^2|x) = \sigma^2 = E(u^2) \tag{43}$$

Now, under the first, fourth and fifth assumptions, we can write:

$$\begin{aligned} E(y|x) &= \beta_0 + \beta_1 x \\ V(y|x) &= \sigma^2 \end{aligned} \tag{44}$$

So the average, or expected, value of y is allowed to change with x, but the variance does not change with x. The constant variance assumption may not be realistic; it must be determined on a case-by-case basis.

Theorem: Sampling variance of OLS. Under assumptions 1 and 2,

we get:

$$\begin{aligned} V(\widehat{\beta_1}|x) &= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ &= \frac{\sigma^2}{SST_x} \quad (45) \\ V(\widehat{\beta_0}|x) &= \frac{\sigma^2(\frac{1}{n}\sum_{i=1}^{n} x_i^2)}{SST_x} \quad (46) \end{aligned}$$

To show this, write, as before,

$$\widehat{\beta_1} = \beta_1 + \sum_{i=1}^{n} w_i u_i \quad (47)$$

where $w_i = \frac{(x_i - \overline{x})}{SST_x}$. We are treating this as nonrandom in the derivation. Because β_1 is a constant, it does not affect $V(\widehat{\beta_1})$. Now, we need to use the fact that, for uncorrelated random variables, the variance of the sum is the sum of the variances. The $\{u_i : i = 1, \ldots, n\}$ are actually independent across i and are uncorrelated. Remember: if we know x, we know w. So:

$$\begin{aligned} V(\widehat{\beta_1}|x) &= Var(\beta_1 + \sum_{i=1}^{n} w_i u_i|x) \quad (48) \\ &= Var\left(\sum_{i=1}^{n} w_i u_i|x\right) \quad (49) \\ &= \sum_{i=1}^{n} Var(w_i u_i|x) \quad (50) \\ &= \sum_{i=1}^{n} w_i^2 Var(u_i|x) \quad (51) \\ &= \sum_{i=1}^{n} w_i^2 \sigma^2 \quad (52) \\ &= \sigma^2 \sum_{i=1}^{n} w_i^2 \quad (53) \end{aligned}$$

where the penultimate equality condition used the fifth assumption so that the variance of u_i does not depend on x_i. Now we have:

$$\begin{aligned} \sum_{i=1}^{n} w_i^2 &= \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{SST_x^2} \quad (54) \\ &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{SST_x^2} \quad (55) \\ &= \frac{SST_x}{SST_x^2} \quad (56) \\ &= \frac{1}{SST_x} \quad (57) \end{aligned}$$

We have shown:

$$V(\widehat{\beta_1}) = \frac{\sigma^2}{SST_x} \tag{58}$$

A couple of points. First, this is the "standard" formula for the variance of the OLS slope estimator. It is *not* valid if the fifth assumption ("homoskedastic errors") doesn't hold. The homoskedasticity assumption is needed, in other words, to derive this standard formula. But, the homoskedasticity assumption is *not* used to show unbiasedness of the OLS estimators. That requires only the first four assumptions we discussed.

Usually, we are interested in β_1. We can easily study the two factors that affect its variance: the numerator and the denominator.

$$V(\widehat{\beta_1}) = \frac{\sigma^2}{SST_x} \tag{59}$$

As the error variance increases – that is, as σ^2 increases – so does the variance in our estimator. The more "noise" in the relationship between y and x (i.e., the larger the variability in u) – the harder it is to learn something about β_1. By contrast, more variation in $\{x_i\}$ is a *good* thing. As $SST_x \uparrow$, $V(\widehat{\beta_1}) \downarrow$.

Notice that $\frac{SST_x}{n}$ is the sample variance in x. We can think of this as getting close to the population variance of x, σ_x^2, as n gets large. This means:

$$SST_x \approx n\sigma_x^2 \tag{60}$$

which means that as n grows, $V(\widehat{\beta_1})$ shrinks at the rate of $\frac{1}{n}$. This is why more data is a good thing – because it shrinks the sampling variance of our estimators.

The standard deviation of $\widehat{\beta_1}$ is the square root of the variance. So:

$$sd(\widehat{\beta_1}) = \frac{\sigma}{\sqrt{SST_x}} \tag{61}$$

This turns out to be the measure of variation that appears in confidence intervals and test statistics.

Next we look at estimating the error variance. In the formula, $V(\widehat{\beta_1}) = \frac{\sigma^2}{SST_x}$, we can compute SST_x from $\{x_i : i = 1, \ldots, n\}$. But we need to estimate σ^2. Recall that $\sigma^2 = E(u^2)$. Therefore, if we could observe a sample on the errors, $\{u_i : i = 1, \ldots, n\}$, an unbiased estimator of σ^2 would be the sample average:

$$\frac{1}{n}\sum_{i=1}^{n} u_i^2 \tag{62}$$

But this isn't an estimator that we can compute from the data we observe because u_i are unobserved. How about replacing each u_i

with its "estimate", the OLS residual $\widehat{u_i}$:

$$u_i = y_i - \beta_0 - \beta_1 x_i \tag{63}$$
$$\widehat{u_i} = y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i \tag{64}$$

Whereas u_i *cannot* be computed, $\widehat{u_i}$ *can be computed* from the data because it depends on the estimators, $\widehat{\beta_0}$ and $\widehat{\beta_1}$. But, except by fluke, $u_i \neq \widehat{u_i}$ for any i.

$$\widehat{u_i} = y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i \tag{65}$$
$$= (\beta_0 + \beta_1 x_i + u_i) - \widehat{\beta_0} - \widehat{\beta_1} x_i \tag{66}$$
$$= u_i - (\widehat{\beta_0} - \beta_0) - (\widehat{\beta_1} - \beta_1) x_i \tag{67}$$

Note that $E(\widehat{\beta_0}) = \beta_0$ and $E(\widehat{\beta_1}) = \beta_1$, but the estimators almost always differ from the population values in a sample. So what about this as an estimator of σ^2?

$$\frac{1}{n} \sum_{i=1}^{n} \widehat{u_i}^2 = \frac{1}{n} SSR \tag{68}$$

It is a true estimator and easily computed from the data after OLS. As it turns out, this estimator is slightly biased: its expected value is a little less than σ^2. The estimator does not account for the two restrictions on the residuals used to obtain $\widehat{\beta_0}$ and $\widehat{\beta_1}$:

$$\sum_{i=1}^{n} \widehat{u_i} = 0 \tag{69}$$
$$\sum_{i=1}^{n} x_i \widehat{u_i} = 0 \tag{70}$$

There is no such restriction on the unobserved errors. The unbiased estimator, therefore, of σ^2 uses a **degrees of freedom** adjustment. The residuals have only $n - 2$ degrees-of-freedom, not n. Therefore:

$$\widehat{\sigma}^2 = \frac{1}{n-2} SSR \tag{71}$$

We now propose the following theorem. **The Unbiased Estimator of σ^2** under the first five assumptions is:

$$E(\widehat{\sigma}^2) = \sigma^2 \tag{72}$$

In regression output, this is the usually reported:

$$\widehat{\sigma} = \sqrt{\widehat{\sigma}^2} \tag{73}$$
$$= \sqrt{\frac{SSR}{(n-2)}} \tag{74}$$

This is an estimator of $sd(u)$, the standard deviation of the population error. One small glitch is that $\widehat{\sigma}$ is not unbiased for σ.[39] This will

[39] There does exist an unbiased estimator of σ but it's tedious and hardly anyone in economics seems to use it. See Holtzman [1950].

not matter for our purposes. $\widehat{\sigma}$ is called the **standard error of the regression**, which means it is an estimate of the standard deviation of the error in the regression. Stata calls it the **root mean squared error**.

Given $\widehat{\sigma}$, we can now estimate $sd(\widehat{\beta_1})$ and $sd(\widehat{\beta_0})$. The estimates of these are called the **standard errors** of the $\widehat{\beta}_j$. We will use these *a lot*. Almost all regression packages report the standard errors in a column next to the coefficient estimates. We just plug $\widehat{\sigma}$ in for σ:

$$se(\widehat{\beta_1}) = \frac{\widehat{\sigma}}{\sqrt{SST_x}} \tag{75}$$

where both the numerator and denominator are computed from the data. For reasons we will see, it is useful to report the standard errors below the corresponding coefficient, usually in parentheses.

Cluster robust standard errors Some phenomena do not affect observations individually, but rather, affect groups of observations which contain individuals. And then it affects those individuals within the group in a common way. Say you wanted to estimate the effect of class size on student achievement, but you know that there exist unobservable things (like the teacher) which affects all the students equally. If we can commit to independence of these unobservables across classes, but individual student unobservables are correlated within a class, then we have a situation where we need to cluster the standard errors. Here's an example:

$$y_{ig} = x'_{ig}\beta + \varepsilon_{ig} \text{ where } 1, \ldots, G$$

and

$$E[\varepsilon_{ig}\varepsilon'_{jg}]$$

which equals zero if $g = g'$ and equals $\sigma_{(ij)g}$ if $g \neq g'$.

Let's stack the data by cluster first.

$$y_g = x'_g\beta + \varepsilon_g$$

The OLS estimator is still $\widehat{\beta} = E[X'X]^{-1}X'Y$. We just stacked the data which doesn't affect the estimator itself. But it does change the variance.

$$V(\beta) = E[[X'X]^{-1}X'\Omega X[X'X]^{-1}]$$

With this in mind, we can now write the variance-covariance matrix for clustered data as

$$\widehat{V}(\widehat{\beta}) = [X'X]^{-1}[\sum_{i=1}^{G} x'_g\widehat{\varepsilon}_g\widehat{\varepsilon}'_g][X'X]^{-1}$$