1. Big Picture and Notation

Context

- An economic model consists of mathematical equations that describe various relationships. Nobel Prize winner Gary Becker postulated a framework to describe an individual's participation in crime:
$$y = f(x_1, x_2, x_3, x_4, \cdots, x_6)$$
where $y =$ hours spend in criminal activity, $x_1 =$ police enforcement, $x_2 =$ hourly wage in legal employment,..., $x_6 =$ age

- After we specify an economic model, we need to turn it into what we call an econometric model. The form of the function $f(\cdot)$ must be specified, and we have to think about how to deal with variables that cannot reasonably be observed, such as "wage" of an hour spent in criminal activity. We can specify the following econometric model for example:
$$crime = \beta_0 + \beta_1 enforcement + \beta_2 wage + \cdots + \beta_6 age + u$$
where $crime =$ some measure of criminal activity, $enforcement =$ some measure of police presence, $wage =$ hourly wage in legal employment. Note that the u term contains the all of the unobserved variables that also explain criminal activity (things like wage for criminal activity, moral character, family background, etc). We will learn how to deal with "unobservables" later in the course.

- The constants β_0, β_1, and β_6 are the parameters of the econometric model, and they describe the directions and strengths of the relationship between crime and the factors used to determine crime in the model.

Regression in the population

- Let's consider a model where crime (y) is only a function of the wage in legal activity (x),
$$y = f(x, u) = \beta_0 + \beta_1 x + u$$
We assume that this is the true data generating process. Note that we have assumed a linear function.

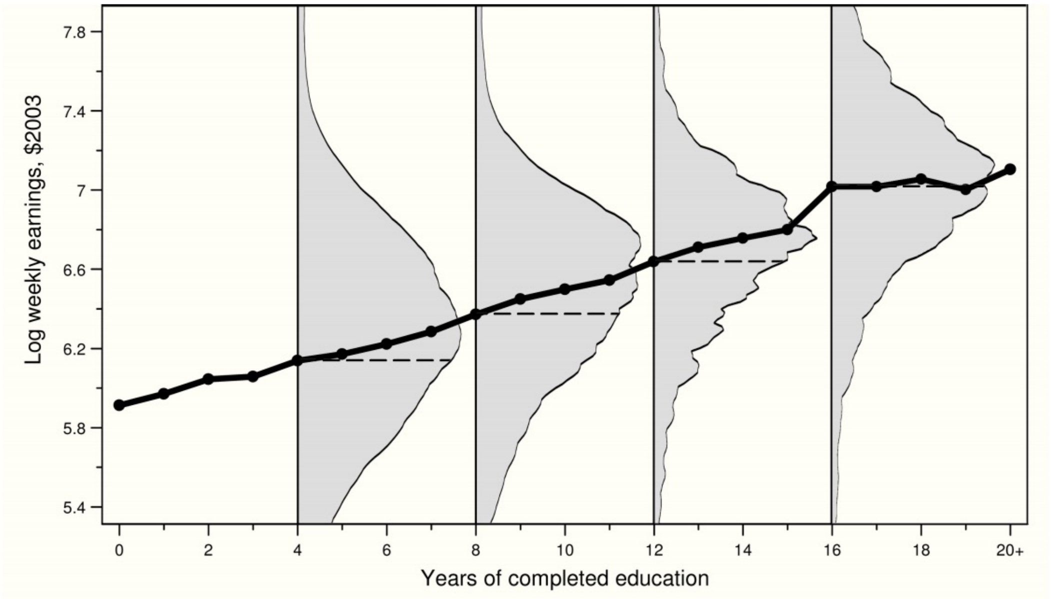
- $u = y - \beta_0 - \beta_1 x$: The variable u is called the error term, or disturbance in the relationship, and represents factors other than x that affect y. The disturbance arises for several reasons, primarily because we cannot hope to capture every influence on an economic variable in a model. We make an important assumption about u that we will revisit later in this section: $E(u|x) = E(u) = 0$

- $E(y|x) = \beta_0 + \beta_1 x$: This is the **population regression function** (PRF). $E(y|x)$ tells us how the population average of one variable changes as we move the conditioning variable over the different values this variable might assume. For every value of the conditioning variable, we get a potentially different average of the dependent variable Y. The collection of all such averages is called the population regression function. In this class, we will assume the PRF is a linear function of x. The linearity implies that a one-unit increase in x changes the *expected* value of y by the amount β_1. For any given value of x, the distribution of y is centered about $E(y|x)$.

Note this definition relies on the assumption (which we will investigate later) that $E(u|x) = E(u)$, which is essentially saying that the value of x doesn't convey any information on average about the value of the disturbance.[1] We can also write:

$$y_i = E(y|x) + u_i$$

This says that any variable y_i can be decomposed into a piece that is explained by x, $E(y|x)$, and some piece that is left over u, which we don't observe.



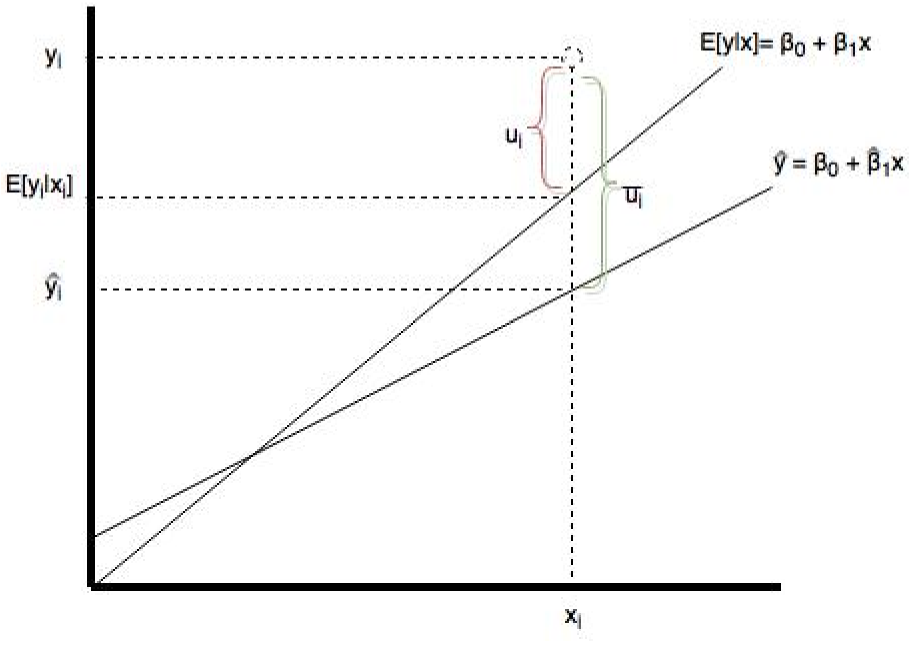
The figure above plots the population regression function of log weekly wages given schooling for men from the 1980 US census. The distribution of earnings is also plotted for a few key values: $4, 8, 12$, and 16 years of schooling.

- The PRF tells us how the average value of y changes with x: it does not say that y equals $\beta_0 + \beta_1 x$ for all units in the population. For example, suppose x=4, then on average this implies log weekly earnings of 6.1 dollars. This does not mean that everyone with 4 years of schooling makes 6.1 dollars.
- In this picture the PRF isn't actually linear, but for the purposes of this class we often assume that it is.

[1] This assumption will allow us to interpret the β coefficient (in the population) as the causal effect of an additional unit of x on the expected value of y. We can still fit a line to our data without this assumption, but we won't be able to interpret the estimate as causal. More on this later, but think of investigating the impact of education on income. If there is something unobservable like ability that varies with the level of education (higher educated people also have more ability) such that $E[u|x] \neq 0$, then we wont be able to say that the coefficient associated with education reveals the true effect of education on income because we are confounding the effect of ability and education

Regression in a sample

- As we have mentioned, we often work with samples rather than the entire population of data. In this case, we do our best to approximate this population regression function.
- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$: This is the fitted regression line. It can be thought of as our best guess for y given a certain value of x. This equation is also called the **sample regression function** (SRF) because it is the estimated version of the PRF.
- $y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\mu} = \hat{y} + \hat{\mu}$: This is now our estimated model. The hat symbol above our beta's indicate that these are calculated estimates of the true beta value they represent. Again we see how we can decompose y_i into two parts: a fitted value (best guess) and a residual.
- $\hat{\mu} = y - \hat{y}$: The variable $\hat{\mu}$ is called the residual, it can be thought of as the deviations between the real y_i value and the predicted $\hat{y}_i$ value.


y_i
$E[y_i|x_i]$
$\hat{y}_i$
u_i
$\bar{u}_i$
$E[y|x]= \beta_0 + \beta_1 x$
$\hat{y} = \beta_0 + \hat{\beta}_1 x$
x_i

The PRF and SRF will (almost) never be the same! But *on average* we will get it right. (*Note*: In the figure, $\bar{u}$ should be $\hat{u}$, and it should be $\hat{\beta}_0$ for the SRF function.)

2. Assumptions of the Linear Regression Model

i. Summary Table

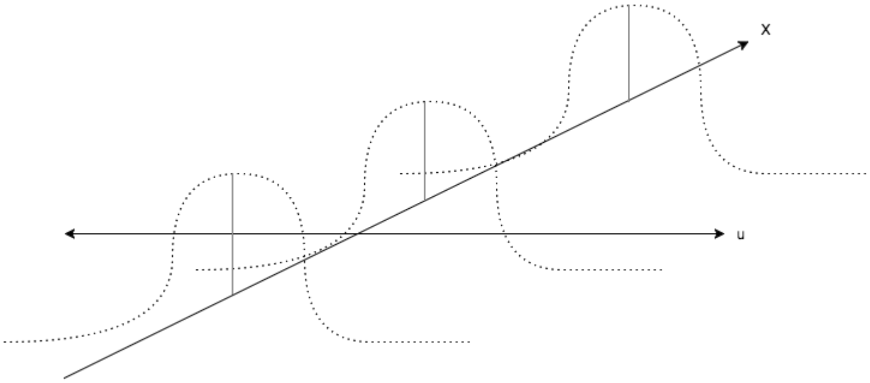
The simple linear regression (SLR) model includes of a set of assumptions about how a data set will be produced by an underlying "data-generating process"

Model	Simple
SLR.1	The population model is linear in parameters $y = \beta_0 + \beta_1 x_1 + u$
SLR.2	$\{(x_i, y_i),\ \ i = 1 \cdots N\}$ is a random sample from the population
SLR.3	The observed explanatory variable (x) is not constant: $Var(x) \neq 0$
SLR.4	No matter what we observe x to be, we expect the unobserved u to be zero $E[u\|x] = 0$
SLR.5	The "error term" has the same variance for any value of x : $Var(u\|x) = \sigma^2$

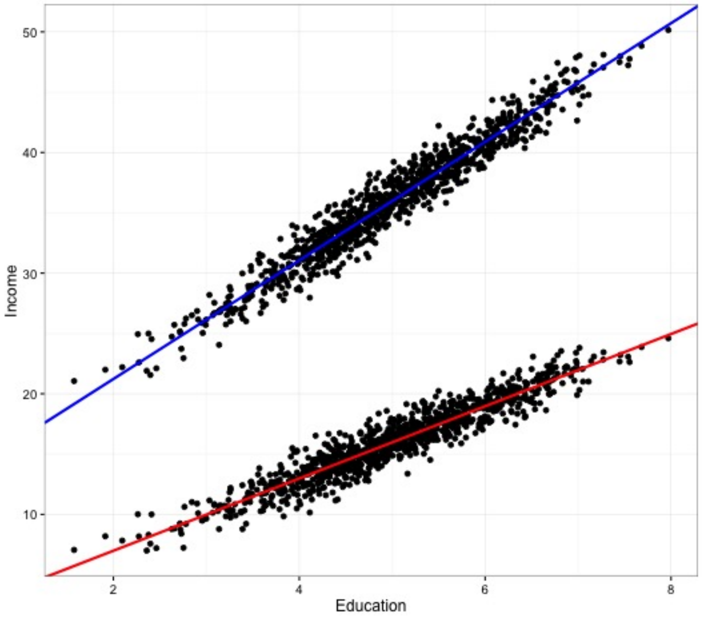
ii. Intuition

1. The assumption of linearity may seem restrictive but since we are referring to the linearity in the *parameters* and the *disturbance*, we are still allowing for the estimation of a variety of nonlinearities (think back to the log-log and log-lin functional forms we saw previously).
2. Random sampling: we want to be able to say something about the population at large. If we obtain information on wages, education, experience, and other characteristics by randomly drawing 500 people from the working population, then we have a random sample from the population of all working people. When is this condition violated? Suppose for example that we are interested in studying factors that influence the accumulation of family wealth. This is a sensitive topic and while we may choose a set of families to interview at random, some families might refuse to report their wealth. If, for example, wealthier families are less likely to disclose their wealth, then the resulting sample on wealth is not a random sample from the population of all families. More on the consequences of this later.
3. We can't estimate the effect of a change in x on y if we don't observe changes in x. If x varies in the population, random samples on x will typically contain variation unless the population variation is minimal or the sample size small. You can check this by taking your data and calculating a few summary statistics, including the variance of x.
4. This is called the Zero Conditional Mean assumption. In words it says that no observations on x convey any information about the expected value of the disturbance. The two graphs below depict this concept:

   - The first graph depicts the condition explicitly $E[u|x = 0]$. It's showing that if you were to plot your error terms (which you never actually see) for every value of x, then you would get a series of distributions centered around 0. It's not as though increases/decreases in the value of x are associated with a positive or negative trends in the expected value of the disturbance.

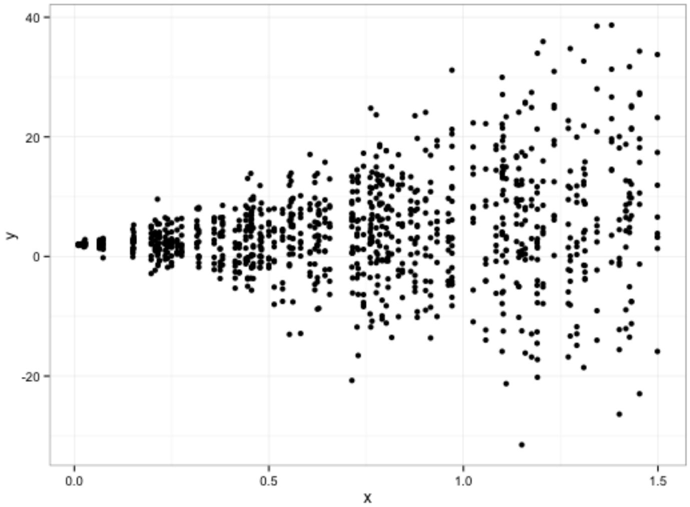


- The second graph tries to give some intuition. Essentially, this assumption will allow us to interpret the β coefficient as the causal effect of an additional unit of x on the expected value of y. We can still fit a line to our data without this assumption, but we won't be able to interpret the estimate as causal. Think of investigating the impact of education on income in the US. If we could account for everything that affected income (ability, education, parent's education, private versus public school attendance) then we could control for all the variables that affect income and isolate the effect of education on it's own $\rightarrow$ red line in the graph. Unfortunately we live in a world where we don't observe everything. Most notably we don't usually observe measures of ability/IQ. If this unobservable characteristic (ability) varies with the level of education (higher educated people also have more ability), and we can't tease the two effects apart, then running a regression of income on education will give us coefficients that CANNOT be interpreted as the causal effect of education $\rightarrow$ blue line in the graph. In other words, we wont be able to say that the coefficient associated with education reveals the true effect of education on income because we are confounding the effect of ability and education.



5. We say that the error term is homoskedastic. Consider a model that describes the profits of firms in an industry as a function of size. Even accounting for size, the profits of large firms will

exhibit greater variation than those of smaller firms. The homoskedasticity assumption would be inappropriate here. If we plot the data and see a fanning out shape (depicted below) this is reason to believe our data does not satisfy this assumption. Note this assumption pertains to the **variance** of the error terms, while SLR4 pertains to the **expected value** of our errors.



3. Properties of $\hat{\beta}_0$, $\hat{\beta}_1$

i. Deriving the Estimators

In lecture, we considered the model $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, and we were given the following formulas for computing $\hat{\beta}_0$ and $\hat{\beta}_1$, which you will use on your problem set:

$$\hat{\beta}_1 = \frac{s_{xy}(x,y)}{s_x^2} = \frac{cov(x,y)}{var(x)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Where did these formulas come from? One way to derive them is to recognize the fact that we want our regression line to minimize the distance between the observed y value and the predicted value $\hat{y}$. In other words we want to to make the set of residuals we obtain very small. There is some debate as to how to go about doing this (some advocate minimizing the absolute value of the residuals, while others argue for minimizing the sum of squared residuals). Minimizing the sum of squared residuals gives more weight to large residuals, that is, outliers in which predicted values are far from actual observation (think of a line of best fit that is trying to "accommodate" the large outliers). More importantly for our purposes, this approach will produce an estimator with desirable properties (more on this later). This is what we refer to as **OLS**. Note we would never choose our estimates to minimize, say, the sum of residuals themselves, as residuals large in magnitude but with opposite signs would tend to cancel out.

In the examples from class and the problem set, you were asked to calculate each term $(x - \bar{x})$, $(y - \bar{y})$ and plug in to the formula to compute $\hat{\beta}_1$ and then $\hat{\beta}_0$.

Question: Why are we using $\bar{x}$ rather than $E(x)$? Answer: because we only have the sample of values we drew, and not the entire population.

Derivation

Let's define W as follows, plugging in our model for $\hat{y}_i$:

$$W = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

We'd like to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ so that W is as small as possible. To do this we solve the following minimization problem with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\min_{\hat{\beta}_0,\hat{\beta}_1} W = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Taking the first order conditions (partial derivatives):

$$\frac{\partial W}{\partial \hat{\beta}_0} = -\sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{1}$$

$$\frac{\partial W}{\partial \hat{\beta}_1} = -\sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0 \tag{2}$$

These equations can be solved for $\hat{\beta}_0$ and $\hat{\beta}_1$ to obtain the expressions for β_0 and β_1 that we saw above. A detailed proof is provided in the appendix.

ii. Properties of $\hat{\beta}$

There are two notable properties of $\hat{\beta}$:

1. $E(u) = 0$: this is simply an assumption about the distribution of the unobservables *in the population*. Without loss of generality, we can assume that things such as average ability are zero in the population of all working people.[2]

2. $E(u|x) = E(u) = 0$: this is the conditional mean assumption. Taking the example in class (which assumed that $u = ability$), this just says that the average ability of individuals in the population is the same regardless of the years of education.

iii. Interpreting estimates: Sign, Size, Significance

Whenever we ask you to "interpret" your estimated results, you need to address these three characteristics of the each coefficient you are examining:

1. **Sign:**

[2] Mathematically, we can always redefine the intercept in to make $E(u) = 0$ true: in the equation $y = \beta_0 + \beta_1 x + u$ where $E(u) \neq 0$ you can add and subtract α from the right and left hand side to get $y = \beta_0 + \alpha + \beta_1 x + (u - \alpha)$. If we call the new error $\varepsilon = u - \alpha_0$, and let $\alpha_0 = E(u)$, then $E(\varepsilon) = 0$

- What sign did you expect the estimated parameter to have? Why? Does your estimate have this sign (i.e. are you surprised or reassured by your results)?

2. **Size:**

   - How do changes in this variable affect the dependent variable according to your estimation? Is this an economically meaningful effect size?

3. **Significance:**

   - Is the estimate statistically different from zero? What is the t-statistic of this hypothesis?

iv. Example: Exercise 2.4 Wooldridge

We have a data set containing information on births to women in the United States. Two variables of interest are the dependent variable, infant birth weight in ounces ($bwght$), and an explanatory variable, average number of cigarettes the mother smoked per day during pregnancy ($cigs$). The following simple regression was estimated using data on 1,388 births:

$$\widehat{bwght} = 119.77 - 0.514 cigs$$

1. Interpret the coefficient on *cigs*
   Sol. Sign: The coefficient on *cigs* is negative, as we would expect since smoking is harmful and will decrease birthweight. Size: Smoking an additional cigarette per day is associated with a 0.514 ounce decrease in predicted birth weight (or you might say that if smoking increases by 1 cigarette, the model *predicts* that birthweight decreases by 0.514 ounces)

2. What is the predicted birth weight when $cigs = 0$? Interpret the intercept term
   Sol. When $cigs = 0$, predicted birth weight is 119.77 ounces. The intercept gives us the predicted birthweight among the sample of women in our dataset who do not smoke. More generally, the intercept is the predicted value of y when $x = 0$, although in some cases it will not make sense to set $x = 0$. In those situations, the intercept is not, in itself, very interesting.

3. To predict a birth weight of 125 ounces, what would *cigs* have to be? Comment.
   Sol. If we want a predicted *bwght* of 125, then

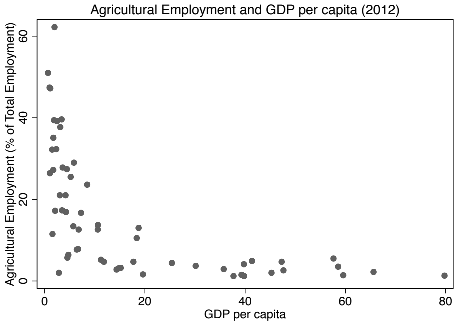
   $$cigs = (125 - 119.77)/(-0.524) \approx -10$$

   This doesn't make sense and illustrates the dangers of trying to predict something as complicated as birth weight with only a single explanatory variable. The largest *predicted* birth weight is necessarily 119.77. Yet almost 700 of the births in the sample had a birth weight higher than 119.77

4. Goodness of Fit, R^2

R^2 is a measure of the "goodness of fit," or how well our regression line fits the data, so that we are able to evaluate the quality of the model after estimating it. Specifically, R^2 is the proportion of variation in our dependent variable, y, that is explained by our model, $\hat{\beta}_0 + \hat{\beta}_1 x$.

Why is the R^2 important? Consider the following data from the World Bank that has this singular shape but we apply the (untransformed) linear model to it anyway. The line that minimizes the sum of squared errors is a flat line even though this clearly misses the underlying relationship between the variables. How can we tell that this is a poor model without looking at it? By looking at the R^2.



Before giving the explicit formula for the R^2, let's define a few additional terms. Consider these three variances, which we will call the Sum of Squares Total (SST), the Sum of Squares Explained (SSE), and the Sum of Squares Residual (SSR).

$$\begin{aligned} SST &= \textstyle\sum_i^n (y_i - \bar{y})^2 \\ SSE &= \textstyle\sum_i^n (\hat{y}_i - \bar{y})^2 \\ SSR &= \textstyle\sum_i^n (y_i - \hat{y})^2 \end{aligned}$$

SST is a measure of the total sample variation in the y_i, that is it measures how spread out the y_i are in the sample. Similarly SSE measures the sample variation in the $\hat{y}_i$ (where we can use the fact $\bar{\hat{y}} = \bar{y}$). Finally the SSR measures the sample variation in the residuals $\hat{u}_i$. The total variation in y can always be expressed as the sum of the explained variation SSE and the unexplained variation SSR. To see this, recall that $y_i = \hat{y}_i + \hat{u}_i$, i.e. the observed value of y_i is equal to the predicted value $\hat{y}_i$ and the difference between the two (the residual) $\hat{u}_i$. The formal proof is in Wooldridge p.39. Thus we can write,

$$SST = SSE + SSR$$

Next, let's define the R^2. We want the R^2 to express how well the regression line fits the data. One way to go about this is to express the fraction of the sample variation that is explained by x (i.e the proportion of variation that is explained by our model). This is reasonable in the sense that if we have a good model, then the sample variation in y should be mostly explained by x (and not by the residual that we don't explicitly observe).

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

The R^2 is always less than 1. Why? Well it's important to understand that if our variables x and y have *some* kind of relationship, knowing what x_i is should give us a little more information about what y is. Ex: if I know nothing about an individual but had to determine the probability that he/she had lung cancer, I would guess the average. However, if I find out that this particular individual is a smoker, I might think there's a slightly higher than average probability that he/she has cancer. We should think of $\hat{y}_i$ as a more knowledgeable guess for y_i than $\bar{y}$ as it uses x_i. Thus, the difference between what we observe and what we predict (SSR) should be smaller than the difference between what we observe and the average (SST). Thus $SSR/SST < 1$ and the $R^2 = 1 - SSR/SST$ will also be less than 1. If the model provides a perfect fit to the data, the $R^2 = 1$.

Example. Wooldridge exercise 2.6: Using data from 1988 for houses sold in Massachusetts, the following equation relates housing prices (*price*) to the distance from a recently built garbage incinerator (*dist*):

$$\widehat{\log(price)} = 9.40 + 0.312 \log(dist)$$
$$n = 135, \;\; R^2 = 0.162$$

a) Interpret the coefficient on $log(dist)$. Is the sign of this estimate what you expect it to be?
Sol. A 10 percent increase in the distance from a recently built garbage incinerator is associated with a 3.12 percent increase in predicted housing prices. The coefficient is positive, as we would expect: if living closer to an incinerator depresses housing prices, then being farther away increases housing prices.

b) How much of the variation in price is explained by the distance to the garbage incinerator?
Sol. We have to use the R-squared to determine how much of the variation in price is explained by the distance to the garbage incinerator. Distance to the incinerator explains only about 16.2 percent of the variation in prices for this sample of houses sold in MA. That means that 83.8 percent of the price variation for these houses is left unexplained. This lack of explanatory power may not be too surprising because many other characteristics should influence housing prices. In a simple regression analysis (with only one explanatory variable), we aren't controlling for all of these other variables explicitly in the model, and hence they are necessarily included in the errors.

c) What other factors about a house affect its price?
Sol. Size of the house, number of bathrooms, size of the lot, age of the home, and quality of the neighborhood (including school quality). In the future we will want to include these explicitly in our econometric model.

1 Appendix: Proofs

Recall that taking derivatives above gave us:

$$\begin{aligned}
\frac{\partial W}{\partial \hat{\beta}_0} &= -\sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\
\frac{\partial W}{\partial \hat{\beta}_1} &= -\sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0
\end{aligned}$$

Beginning with the first equation (eq 1 in the body of the notes):

$$\begin{aligned}
2\left[-\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\beta}_1 x_i\right] &= 0 && \text{Distribute the Summation} \\
\left[-\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\beta}_1 x_i\right] &= 0 && \text{Get rid of the 2} \\
\sum_{i=1}^{n} \hat{\beta}_0 = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} &\hat{\beta}_1 x_i && \text{Re-arranging terms}
\end{aligned}$$

Since β_0 and β_1 are same for all cases in the original linear equation, this further simplifies to:

$$\begin{aligned}
n\hat{\beta}_0 &= \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i \\
\beta_0 &= \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i \\
\beta_0 &= \bar{y} - \hat{\beta}_1 \bar{x}
\end{aligned}$$

This is our final expression for β_0. Going back to equation (2) we will solve for β_1:

$$\begin{aligned}
-\sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i &= 0 \\
\sum_{i=1}^{n} 2x_i(-y_i + \hat{\beta}_0 + \hat{\beta}_1 x_i) &= 0 \\
\sum_{i=1}^{n} x_i(-y_i + \hat{\beta}_0 + \hat{\beta}_1 x_i) &= 0 \\
\sum_{i=1}^{n} x_i(-y_i + (\bar{y} - \hat{\beta}_1\bar{x}) + \hat{\beta}_1 x_i) &= 0 \\
\sum_{i=1}^{n} x_i(-y_i + (\bar{y} - \hat{\beta}_1\bar{x}) + \hat{\beta}_1 x_i) &= 0 \\
\sum_{i=1}^{n} x_i(\bar{y} - y_i + \hat{\beta}_1(x_i - \bar{x})) &= 0 \\
\sum_{i=1}^{n} x_i(\hat{\beta}_1(x_i - \bar{x})) &= \sum_{i=1}^{n} x_i(y_i - \bar{y}) \\
\hat{\beta}_1 \sum_{i=1}^{n} x_i(x_i - \bar{x}) &= \sum_{i=1}^{n} x_i(y_i - \bar{y})
\end{aligned}$$

From some properties of summation operation (see Appendix A.1 Wooldridge for the full set of steps)

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

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{2}$$

Then plugging (1) and (2) into our previous expression:

$$\begin{aligned}
\hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}
\end{aligned}$$