1. Population Parameters and Sample Estimators[1]

i. Difference between Sample and Population[2]

Typically, we do not have the full set of population data to work with. This is why we take samples. We can use random samples to infer something about the population at large. E.g., we use the sample mean as a close approximation to the population mean.

Random Sample

Formally, let X be a random variable representing a population with a probability density function f_X.

Definition: If $X_1, X_2 \cdots, X_n$ are independent random variables with a common probability density function, then $\{X_1, \cdots X_n\}$ is said to be a random sample from the population represented by that same PDF. The random nature of $X_1, X_2 \cdots, X_n$ in the definition of random sampling illustrates that many different outcomes are possible before the sampling is actually carried out.

Example: You decide you want to investigate the effect of education on family income. If you obtain data on family income from a sample of n=100 families in the US, the incomes you observe will usually differ for each different sample of 100 families. Once a sample is obtained, you will have a set of numbers, which we denote $\{x_1, x_2, \cdots, x_n\}$

Estimator

Given a random sample $X_1, X_2 \cdots, X_n$ drawn from a PDF that depends on some unknown parameter θ, an estimator of θ is a rule that assigns each possible outcome of the sample a value of $\hat{\theta}$. E.g., if the population has mean μ, then an **estimator** of μ is a **rule (function)** that assigns each possible outcome of the sample a value of $\hat{\mu}$. The rule is specified before any sampling is carried out. The sample mean $\bar{X}$ is an estimator for the population mean μ, and the sample variance s_X^2 is an estimator for the population variance σ_X^2

In words: if X is a variable generated by a random process, such as throwing a die, then $E[X] = \mu_X$ is the average in infinitely many repetitions of this process. If X is a variable that comes from a survey, $E[X]$ is the average obtained if everyone in the population from which the observation is drawn were to be surveyed. We usually do not have infinite repetitions, or the ability to survey the entire population, therefore we calculate the average in a sample $\bar{X}$, and say this is the best *estimate* for the average in the population.

Important to remember: $\bar{X}, \bar{Y}$ are random variables because they depend on the random sample: as we obtain different random samples from the population, the values of $\bar{X}, \bar{Y}$ can change. Hence they have a certain probability distribution, with a certain mean and a certain

[1] Many thanks to Erin Kelley for initially drafting these notes.
[2] This section is quoted from Wooldridge p.748-750

variance.

$$E[\bar{X}] = E[\frac{1}{n}\sum_i X_i] = \frac{1}{n}E[\sum_i X_i] = \frac{1}{n}nE[X_i] == \frac{1}{n}n(\mu) = \mu_X$$

$$Var[\bar{X}] = Var[\frac{1}{n}\sum_i X_i] = \frac{1}{n^2}Var[\sum_i X_i] = \frac{1}{n^2}nVar[X_i] = \frac{1}{n^2}n\sigma_X^2 = \frac{\sigma_X^2}{n}$$

$$Sd[\bar{X}] = \sqrt{(Var[\bar{X}])} = \frac{\sigma_X}{\sqrt{n}}$$

A few points:

- If we repeat our random draws many times, or we survey more and more people, then the sample average will approach the population average. We talked about this in an earlier lecture and section.
- To actually calculate the variance and standard deviation of $\bar{X}$, we need to know σ_X^2. However, as our table shows us, these terms can only be calculated from the entire population. Lucky for us, we have an **estimator** for σ_X^2, which is s_X^2. As a matter of terminology, an estimate for the standard deviation of a variable, is referred to as the **standard error**.

Finally, once a sample is obtained, we have a set of numbers say, $\{x_1, x_2, \cdots, x_n\}$, which constitute the data that we work with. For these actual data outcomes $\{x_1, x_2, \cdots, x_n\}$, the **estimate**, $\bar{x}$, is just the average in the sample:

$$\bar{x} = \frac{1}{n}\sum_i^n x_i = \frac{(x_1 + x_2 \cdots + x_n)}{n}$$

Summary: X as a continuous variable

	Symbol	Formula
Population parameters	μ	$\sum_{j=1}^k x_j f(x_j)$
	σ_X^2	$E[(X - E(X))^2]$
	σ_X	$\sqrt{E[(X - E(X))^2]}$
Sample estimators	$\bar{X}$	$\frac{1}{n}\sum_i X_i$
	s_X^2	$\frac{1}{n-1}\sum_i (X_i - \bar{X})^2$
	s_X	$\sqrt{\frac{1}{n-1}\sum_i (X_i - \bar{X})^2}$
Estimator parameters	$E(\bar{X})$	μ_X
	$Var(\bar{X})$	$\frac{\sigma_X^2}{n}$
	$Sd(\bar{X})$	$\frac{\sigma_X}{\sqrt{n}}$
SE of estimator	$Se(\bar{X})$	$\frac{s_X}{\sqrt{n}}$

Estimators and OLS:

Now, transitioning back to the population model we discussed previously:

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

$\hat{\beta}_0$ and $\hat{\beta}_1$ are **estimators** for the parameters β_0 and β_1. Indeed we derived a formula for our $\hat{\beta}'s$, this was a rule that assigns each possible outcome of the sample a value of β. Then, for the given sample of data we work with, we obtain intercept and slope **estimates**, $\hat{\beta}_0$ and $\hat{\beta}_1$. Using $\hat{\beta}_0$ and $\hat{\beta}_1$ we can calculate the fitted values $\hat{y}_i$ for each observation from the equation.

Finally, if we calculate the standard deviation associated with our estimator, we get the **standard error**. Why do our $\hat{\beta}$ estimates have a variance/standard deviation? Because in theory we can draw many different random samples and compute the $\hat{\beta}$ estimates for each one. This would yield a distribution of our $\hat{\beta}$ estimates, with a mean and a variance! In practice however we will use a formula for the variance of our $\hat{\beta}$ which depends on the variation in our error terms, and the variation in our x variables. With more unexplained variation in our data, our estimate of the relationship between x on y (i.e. our estimate of $\hat{\beta}$) will be more variable as well (i.e. it $var(\hat{\beta})$ will be higher). R (or other computational software) uses this formula in the background as well.

Next we will go through some assumptions about the properties of these estimators, which will allow us to calculate their associated mean and variance.

2. Assumptions of the Linear Regression Model

i. Summary Table

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

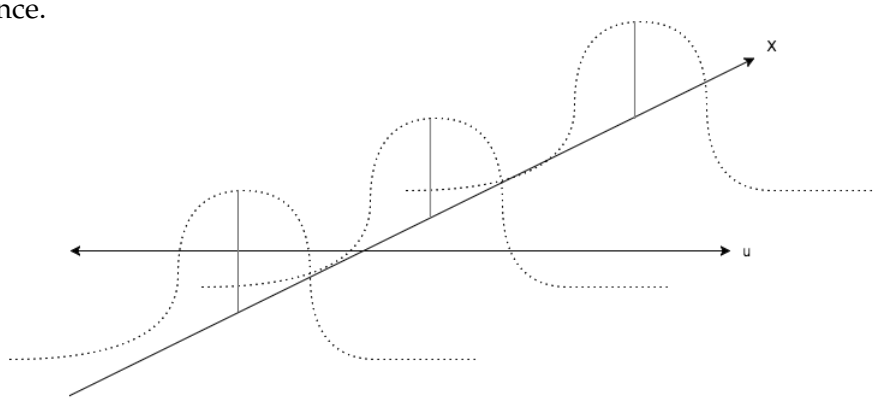
Model	Simple
SLR.1	The population model is linear in parameters $y = \beta_0 + \beta_1 x_1 + u$
SLR.2	$\{(x_i, y_i), \ \ i = 1 \cdots N\}$ is a random sample from the population
SLR.3	The observed explanatory variable (x) is not constant: $Var(x) \neq 0$
SLR.4	No matter what we observe x to be, we expect 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

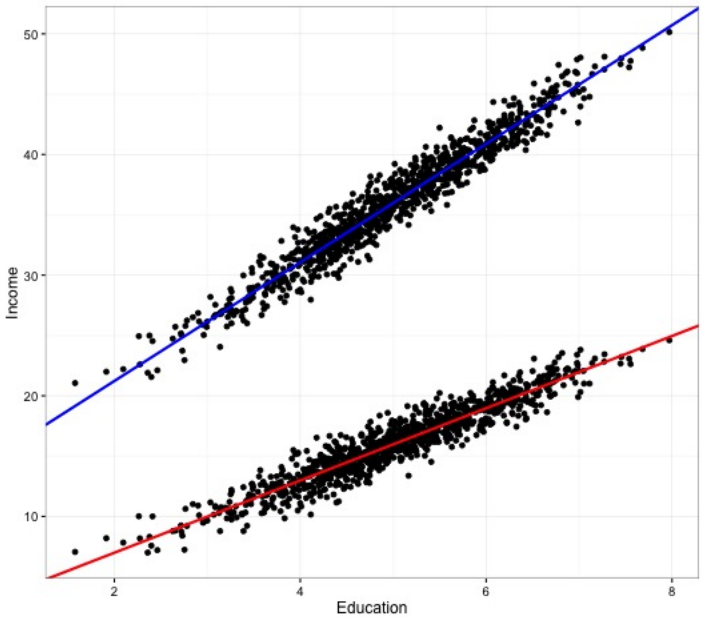
1. The assumption of linearity may seem restrictive but since we are referring to the linearity in the *parameters* and the *disturbance*, we are still allowing for the estimation of a variety of nonlinearities (think back to the log-log and log-lin functional forms we saw previously).

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 chose 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. 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 depicts this concept:

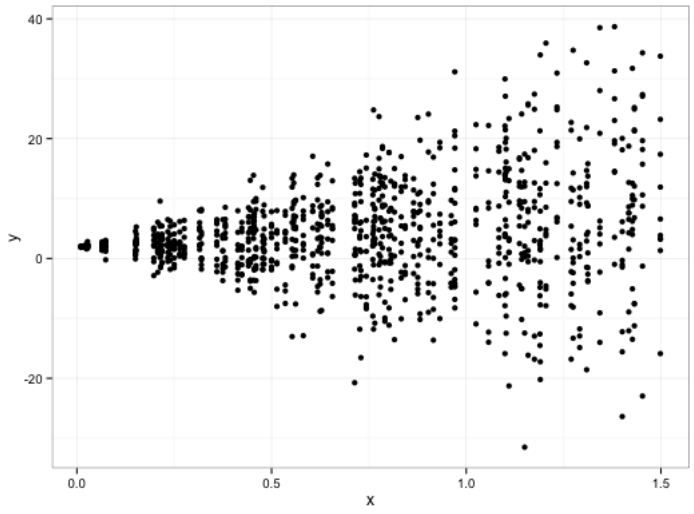
   - The first graph depicts the condition explicitly $E[u|x = 0]$. It's showing that if you were to plot your error terms (which you never actually see) for every value of x, then you would get a series of distributions centered around 0. It's not as though increases/decreases in the value of x are associated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



iii. Implications of Assumptions

From these assumptions we get: [3]

1. $E(\hat{\beta}_1) = \beta_1$

Proof:

$$\begin{aligned}
E(\hat{\beta}_1) &= E\left[\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \\
&= E\left[\frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x}) x_i}\right] \quad \text{See Appendix A Woolridge} \\
&= E\left[\frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \mu_i)}{\sum_{i=1}^n (x_i - \bar{x}) x_i}\right] \\
&= \beta_0 \cdot E\left[\frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x}) x_i}\right] + \beta_1 \cdot E\left[\frac{\sum_{i=1}^n (x_i - \bar{x}) x_i}{\sum_{i=1}^n (x_i - \bar{x}) x_i}\right] + E\left[\frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{\sum_{i=1}^n (x_i - \bar{x}) x_i}\right] \\
&= \beta_1 \cdot E\left[\frac{\sum_{i=1}^n (x_i - \bar{x}) x_i}{\sum_{i=1}^n (x_i - \bar{x}) x_i}\right] + E\left[\frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{\sum_{i=1}^n (x_i - \bar{x}) x_i}\right] \quad \text{Recalling that } \sum_{i=1}^n (x_i - \bar{x}) = 0 \\
&= \beta_1 + E\left[\frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{\sum_{i=1}^n (x_i - \bar{x}) x_i}\right] \quad \text{Canceling} \\
&= \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x}) x_i} E\left[u_i | x_1, \cdots x_n\right] \\
&= \beta_1 \qquad \text{SLR2 and SLR4}
\end{aligned}$$

2. $E(\hat{\beta}_0) = \beta_0$

Proof:

$$\begin{aligned}
E(\hat{\beta}_0) &= E[\bar{y} - \hat{\beta}_1 \bar{x}] \\
&= E[(\beta_0 + \beta_1 \bar{x} + \bar{u}) - \hat{\beta}_1 \bar{x}] \quad \text{Plug in for } \bar{y} \\
&= E[(\beta_0 + \bar{x}(\beta_1 - \hat{\beta}_1) + \bar{u}] \\
&= \beta_0 + E[\bar{x}(\beta_1 - \hat{\beta}_1)] + E[\bar{u} | x_1, \cdots x_n] \\
&= \beta_0 + E[\bar{x}(\beta_1 - \hat{\beta}_1)] \qquad \text{SLR2 and SLR4} \\
&= \beta_0 \quad \text{Because we showed } E(\hat{\beta}_1) = \beta_1
\end{aligned}$$

[3] In this proof, the expected values are conditional on the sample values of the independent variable x. To avoid excessive notation we keep the conditioning on $\{x_1, x_2, \cdots x_n\}$ implicit. Remember that expressions that are functions only of the x_j are nonrandom in the conditioning and can be pulled out of the expected value

3. $Var(\hat{\beta}_1) = \sigma_u^2 / SST_x$

Proof:

$$\begin{aligned}
var(\hat{\beta}_1|X) &= var\left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}|X\right) \\
&= var\left(\beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})u_i}{\sum_{i=1}^n (x_i - \bar{x})^2}|X\right) \\
&= \left(\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)^2 var\left(\sum_{i=1}^n (x_i - \bar{x})u_i|X\right) \\
&= \frac{1}{SST_x^2} var\left(\sum_{i=1}^n (x_i - \bar{x})u_i|X\right) \\
&= \frac{1}{SST_x^2}\left(\sum_{i=1}^n (x_i - \bar{x})^2\right) var\left(u_i|\, X\right) \\
&= \frac{1}{SST_x^2} SST_x var\left(u_i|\, X\right) \\
&= \frac{\sigma_u^2}{SST_x}
\end{aligned}$$

Where[4]:

$$\begin{aligned}
\sigma_u^2 &= Var(u|x) \\
&= E(u^2|x) - [E(u|x)]^2 \quad \text{Definition of Variance} \\
&= E(u^2|x) \quad \text{SLR4} \\
&= E(u^2)
\end{aligned}$$

So we think to use the following unbiased "estimator" of σ^2

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

But this isn't a true estimator because we do not observe the errors u_i. However, we do have an estimator for these errors, which we refer to as the residuals, denoted $\hat{u}_i$.
Then the estimator of σ_u^2 is given by the rule [5]:

$$\begin{aligned}
\hat{\sigma}^2 &= s^2 \\
&= \frac{1}{n-2}\sum_{i=1}^n \hat{u}_i^2 \\
&= \frac{1}{n-2}\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{SSR}{n-2}
\end{aligned}$$

[4] We can see that the variation in $\hat{\beta}_1$ increases with the greater variance of our errors (more variation in the unobservables affecting y makes it difficult to pin down an estimate for β_1), and decreases with greater variance in our x's (when the data is more spread out, we can more easily trace out a relationship between x and $E[y|x]$. Note more observations increases the variation in x_i

[5] See Woolridge p.57 for a good explanation for why we have to divide by $n-2$ rather than n like we might assume

Then, the standard deviation of our estimator (which if you recall, we say standard error) is:

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}_u^2}{SST_x}} = \frac{\hat{\sigma}_u}{\sqrt{SST_x}}$$

In class we also used the notation $se(\hat{\beta}_1) = \sqrt{\hat{var}(\hat{\beta}_1)}$. Note that $se(\hat{\beta}_1)$ is considered a random variable. Indeed if we run OLS with different samples, we will get different $\hat{\sigma}$ and therefore a different $se(\hat{\beta}_1)$. For a given sample, $se(\hat{\beta}_1)$ is a number, just as $\hat{\beta}_1$ is simply a number when we compute it from the given data.

4. $Var(\hat{\beta}_0)$

Proof:

$$\begin{aligned} Var(\hat{\beta}_0) &= \frac{\sigma^2 n^{-1} \sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sigma^2 n^{-1} \sum_{i=1}^n x_i^2}{SST_x} \end{aligned}$$

iv. Summary table

	Symbol	Formula
Population parameters	β_0	
	β_1	
Sample estimators	$\hat{\beta}_0$	$\bar{y} - \hat{\beta}_1 \bar{x}$
	$\hat{\beta}_1$	$\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$
Estimator parameters*	$E(\hat{\beta}_0)$	β_0
	$E(\hat{\beta}_1)$	β_1
	$Var(\hat{\beta}_1)$	$\frac{\sigma_u^2}{SST_x}$
	$Sd(\hat{\beta}_1)$	$\frac{\sigma_u}{\sqrt{SST_x}}$
SE of estimator	$Se(\hat{\beta}_1)$	$\frac{\hat{\sigma}_u}{\sqrt{SST_x}}$

*We don't show $Var(\hat{\beta}_0)$, $Sd(\hat{\beta}_0)$, or $Se(\hat{\beta}_0)$ because we rarely care.

3. Multiple Linear Regression

i. Motivation

What we saw in the Single Linear regression model follows through. Let's review our terms:

- $E[y|x_1, \cdots x_k]$: This is the population regression function (PRF), $E[y|x_1, \cdots x_k]$ is a linear function of x.
- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$: 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.

Why are we turning to multiple regression analysis?

1. It allows us to explicity control for other variables that could be affecting our dependent variable y
2. By adding more variables into our regression function, we can explain more of the variation in y and better predict the outcomes we are interested in.
3. We can introduce more general functional form relationships

The following two equations illustrate these benefits explicitly:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u \tag{1}$$

$$consumption = \beta + 0 + \beta_1 inc + \beta_2 inc^2 + u \tag{2}$$

In equation (1) we want to know the effect of education on wages. Here we explicitly control for experience. Compared to the single linear regression model, we have effectively taken *experience* out of the error term and put it explicitly in the equation. In a simple regression analysis, we would have had to assume that experience is uncorrelated with education, an unrealistic assumption.

In equation (2) the model falls outside simple regression because it contains two functions of income, *income* and $income^2$ (more on the interpretation of quadratics later)

ii. Assumptions

Model	Multiple
MLR.1	The population model is linear in parameters $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \mu$
MLR.2	$\{(x_{i1}, \cdots, x_{ik}, y_i), \ \ i = 1 \cdots N\}$ is a random sample from the population
MLR.3	No perfect collinearity among observed variables and $Var(x_j) \neq 0, j = 1 \cdots k$
MLR.4	No matter what we observe $(x_{i1}, \cdots, x_{ik})$ to be, we expect the unobserved u to be zero $E[u|x_1, \cdots, x_k] = 0$
MLR.5	The "error term" has the same variance for any value of$(x_1, \cdots x_k) : Var(u|x_1, \cdots x_k) = \sigma^2$

A few words about these assumptions now that we are dealing with x variables:

- According to Woolridge p.85 "Assumption MLR.3 is more complicated than its counterpart for simple regression because we must now look at relationships between all independent variables. If an independent variable is an exact linear combination of the other independent variables, then we say the model suffers from perfect collinearity, and it cannot be estimated by OLS. It is important to note that Assumption MLR.3 does allow the independent variables to be correlated; they just cannot be perfectly correlated."
- Consider the following example to gain a bit more intuition about MLR4. The population regression equation is given by:

  $$income = \beta_0 + \beta_1 educ + \beta_2 exp + u$$

  The assumption here is that $E[u|educ, exper] = 0$. This implies "other factors affecting wage are not related on average to education and experience. Therefore, if we think innate ability is part of u, then we will need average ability levels to be the same across all combinations of education and experience in the working population" (Woolridge p.70).

iii. Implications of these assumption

MLR1-MLR4 gets us the unbiasedness property we want:

$$E[\hat{\beta}_j] = \beta_j$$

MLR5 gets us a formula for the variance of $\hat{\beta}$

$$Var(\hat{\beta}_j) = \frac{\sigma_u^2}{SST_j(1 - R_j^2)}$$

where $SST_j = \sum_i (x_{ij} - \bar{x})^2$ is the total sample variation in x_j, and R_j^2 is the R squared from regressing x_j on all other independent variables. More on this in the next set of lecture notes.

iv. OLS

Derivation

We apply the method of minimizing the sum of squared residuals in this framework as well. Specifically, we will obtain $\hat{\beta}_0 ... \hat{\beta}_1 ... \hat{\beta}_k$ by minimizing:

$$min \quad W = \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} \cdots \hat{\beta}_k x_{ik})^2$$

As mentioned in class, the expressions we get from this derivation are complicated, and we will let R do the work for us.

Interpretation

Take the following OLS regression sample equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 educ + \hat{\beta}_2 exp = 1.27 + 0.543 educ + 0.27 exp$$

First, the intercept $\hat{\beta}_0 = 1.27$ is the predicted y if education and experience are both set to zero. Often this intercept is not by itself very meaningful. The estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ are interpreted as partial effects:

- A one year increase in education is associated with a $\hat{\beta}_1$ increase in predicted wage holding experience constant. In other words, if we choose two individuals, A and B, and these individuals have the same years of work experience, but the education of Person A is one year higher than the education of Person B, then we predict Person A to have a wage that is 0.543 thousands of dollars (of 543 dollars) higher than Person B. (This says nothing about any two actual people, but it is our best prediction.)
- A one year increase in experience is associated with a $\hat{\beta}_2$ increase in predicted wage holding education constant. In other words, if we choose two individuals, A and B, and these individuals have the same years of education , but the work experience of Person A is one year higher than the work experience of Person B, then we predict Person A to have a wage that is 0.270 thousands of dollars (of 270 dollars) higher than Person B. (This says nothing about any two actual people, but it is our best prediction.)

v. Example 3.2 Woolridge

Data on working men was used to estimate the following equation

$$\widehat{educ} = 10.36 - 0.094sibs + 0.131meduc + 0.210feduc$$

where *educ* is years of schooling, *sibs* is number of siblings, *meduc* is mother's years of schooling, and $feduc$ is father's years of schooling

1. Does sibs have the expected effect?

   Having one additional sibling is associated with a 0.094 decrease in predicted years of education, holding *meduc* and $feduc$ fixed. Because of budget constraints, it makes sense that, the more siblings there are in a family, the less education any one child in the family has.
2. Holding meduc and feduc fixed, by how much does sibs have to increase to reduce predicted years of education by one year? To find the increase in the number of siblings that reduces predicted education by one year, we solve:

$$\begin{aligned} 1 &= 0.094(\Delta sibs) \\ \Delta sibs &= \frac{1}{0.094} \\ &= 10.6 \end{aligned}$$

3. Discuss the interpretation of the coefficient on *meduc*

   One more year of mother's education is associated with 0.131 years more of predicted education holding *sibs* and $feduc$ fixed. So if a mother has four more years of education, her son is predicted to have about a half a year (.524) more years of education.
4. Suppose that Man A has no siblings, and his mother and father each have 12 years of education. Man B has no siblings, and his mother and father each have 16 years of education. What is the predicted difference in years of education between B and A?

For Man A we have:

$$\widehat{educ} = 10.36 - 0.094(0) + 0.131(12) + 0.210(12)$$

For man B we have:

$$\widehat{edu} = 10.36 - 0.094(0) + 0.131(16) + 0.210(16)$$

So the predicted difference in education between B and A is

$$0.131(4) + 0.210(4) = 1.364$$

vi. Running MLR in R

Click this text to access notes in Jupyter.