Refresher:

$$y = \beta_0 + \beta_1 x + u. \quad \textbf{(2.1)}$$

The most difficult issue to address is whether model (2.1) really allows us to draw ceteris paribus conclusions about how x affects y. We just saw in equation (2.2) that β_1 *does* measure the effect of x on y, holding all other factors (in u) fixed. Is this the end of the causality issue? Unfortunately, no. How can we hope to learn in general about the ceteris paribus effect of x on y, holding other factors fixed, when we are ignoring all those other factors?

As we will see in Section 2.5, we are only able to get reliable estimators of β_0 and β_1 from a random sample of data when we make an assumption restricting how the unobservable u is related to the explanatory variable x. Without such a restriction, we will not be able to estimate the ceteris paribus effect, β_1. Because u and x are random variables, we need a concept grounded in probability.

Before we state the key assumption about how x and u are related, there is one assumption about u that we can always make. As long as the intercept β_0 is included in the equation, nothing is lost by assuming that the average value of u in the population is zero.

Mathematically,

$$\mathrm{E}(u) = 0. \quad \textbf{(2.5)}$$

Importantly, assume (2.5) says nothing about the relationship between u and x but simply makes a statement about the distribution of the unobservables in the population.

We now turn to the crucial assumption regarding how u and x are related. A natural measure of the association between two random variables is the *correlation coefficient*. (See Appendix B for definition and properties.) If u and x are *uncorrelated*, then, as random variables, they are not *linearly* related. Assuming that u and x are uncorrelated goes a long way toward defining the sense in which u and x should be unrelated in equation (2.1). But it does not go far enough, because correlation measures only linear dependence between u and x. Correlation has a somewhat counterintuitive feature: it is possible for u to be uncorrelated with x while being correlated with functions of x, such as x^2. (See Section B.4 for further discussion.) This possibility is not acceptable for most regression purposes, as it causes problems for interpreting the model and for deriving statistical properties. A better assumption involves the *expected value of u given x*.

Because u and x are random variables, we can define the conditional distribution of u given any value of x. In particular, for any x, we can obtain the expected (or average) value of u for that slice of the population described by the value of x. The crucial assumption is that the average value of u does *not* depend on the value of x. We can write this as

$$\mathrm{E}(u|x) = \mathrm{E}(u) = 0, \quad \textbf{(2.6)}$$

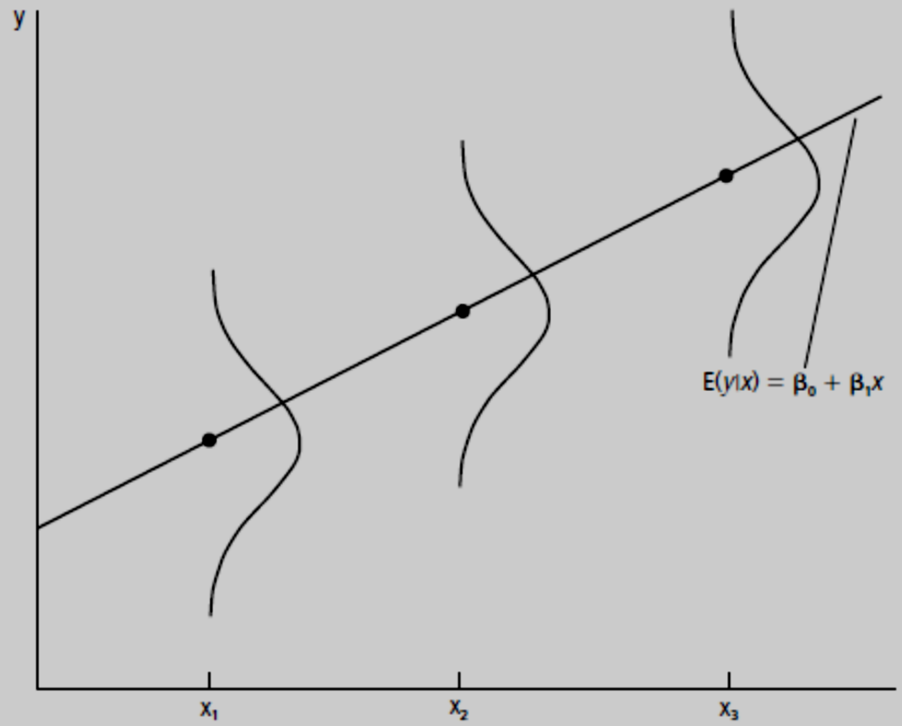
where the second equality follows from (2.5). The first equality in equation (2.6) is the new assumption, called the **zero conditional mean assumption**. It says that, for any given value of x, the average of the unobservables is the same and therefore must equal the average value of u in the entire population.

$$E(y|x) = \beta_0 + \beta_1 x \tag{2.8}$$

Equation (2.8) shows that the **population regression function (PRF)**, $E(y|x)$, is a linear function of x. The linearity means 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 $E(y|x)$, as illustrated in Figure 2.1.

Figure 2.1

$E(y|x)$ as a linear function of x.



There are several ways to motivate the following estimation procedure. We will use (2.5) and an important implication of assumption (2.6): in the population, u has a zero mean and is uncorrelated with x. Therefore, we see that u has zero expected value and that the *covariance* between x and u is zero:

$$E(u) = 0 \tag{2.10}$$

$$\text{Cov}(x,u) = E(xu) = 0, \tag{2.11}$$

where the first equality in (2.11) follows from (2.10). (See Section B.4 for the definition and properties of covariance.) In terms of the observable variables x and y and the unknown parameters β_0 and β_1, equations (2.10) and (2.11) can be written as

$$E(y - \beta_0 - \beta_1 x) = 0 \tag{2.12}$$

and

$$E[x(y - \beta_0 - \beta_1 x)] = 0, \tag{2.13}$$

respectively. Equations (2.12) and (2.13) imply two restrictions on the joint probability distribution of (x,y) in the population. Since there are two unknown parameters to estimate, we might hope that equations (2.12) and (2.13) can be used to obtain good estimators of β_0 and β_1. In fact, they can be. Given a sample of data, we choose estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ to solve the *sample* counterparts of (2.12) and (2.13):

$$n^{-1}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0. \tag{2.14}$$

$$n^{-1}\sum_{i=1}^{n}x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0. \tag{2.15}$$

This is an example of the *method of moments* approach to estimation. (See Section C.4 for a discussion of different estimation approaches.) These equations can be solved for $\hat{\beta}_0$ and $\hat{\beta}_1$.

Using the basic properties of the summation operator from Appendix A, equation (2.14) can be rewritten as

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}, \tag{2.16}$$

where $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$ is the sample average of the y_i and likewise for $\bar{x}$. This equation allows us to write $\hat{\beta}_0$ in terms of $\hat{\beta}_1$, $\bar{y}$, and $\bar{x}$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}. \tag{2.17}$$

Therefore, once we have the slope estimate $\hat{\beta}_1$, it is straightforward to obtain the intercept estimate $\hat{\beta}_0$, given $\bar{y}$ and $\bar{x}$.

Dropping the n^{-1} in (2.15) (since it does not affect the solution) and plugging (2.17) into (2.15) yields

$$\sum_{i=1}^{n} x_i(y_i - (\bar{y} - \hat{\beta}_1\bar{x}) - \hat{\beta}_1 x_i) = 0$$

which, upon rearrangement, gives

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \hat{\beta}_1\sum_{i=1}^{n} x_i(x_i - \bar{x}).$$

From basic properties of the summation operator [see (A.7) and (A.8)],

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

Therefore, provided that

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0, \tag{2.18}$$

the estimated slope is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{2.19}$$

Equation (2.19) is simply the sample covariance between x and y divided by the sample variance of x. (See Appendix C. Dividing both the numerator and the denominator by $n - 1$ changes nothing.) This makes sense because β_1 equals the population covariance divided by the variance of x when $E(u) = 0$ and $Cov(x,u) = 0$. An immediate implication is that if x and y are positively correlated in the sample, then $\hat{\beta}_1$ is positive; if x and y are negatively correlated, then $\hat{\beta}_1$ is negative.

Although the method for obtaining (2.17) and (2.19) is motivated by (2.6), the only assumption needed to compute the estimates for a particular sample is (2.18). This is hardly an assumption at all: (2.18) is true provided the x_i in the sample are not all equal to the same value. If (2.18) fails, then we have either been unlucky in obtaining our sample from the population or we have not specified an interesting problem (x does not vary in the population.). For example, if $y = wage$ and $x = educ$, then (2.18) fails only if everyone in the sample has the same amount of education. (For example, if everyone is a high school graduate. See Figure 2.3.) If just one person has a different amount of education, then (2.18) holds, and the OLS estimates can be computed.

```
. bcuse ceosal1

Contains data from http://fmwww.bc.edu/ec-p/data/wooldridge/ceosal1.dta
  obs:         209
 vars:          12                    25 Sep 2012 14:44
 size:       6,270
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
salary          int    %9.0g                  1990 salary, thousands $
pcsalary        int    %9.0g                   % change salary, 89-90
sales           float  %9.0g                  1990 firm sales, millions $
roe             float  %9.0g                   return on equity, 88-90 avg
pcroe           float  %9.0g                  % change roe, 88-90
ros             int    %9.0g                  return on firm's stock, 88-90
indus           byte   %9.0g                  =1 if industrial firm
finance         byte   %9.0g                  =1 if financial firm
consprod        byte   %9.0g                  =1 if consumer product firm
utility         byte   %9.0g                  =1 if transport. or utilties
lsalary         float  %9.0g                  natural log of salary
lsales          float  %9.0g                  natural log of sales
-------------------------------------------------------------------------------

. reg  salary roe

      Source |       SS       df       MS              Number of obs =     209
-------------+------------------------------           F(  1,   207) =    2.77
       Model |  5166419.04     1  5166419.04           Prob > F      =  0.0978
    Residual |   386566563   207  1867471.32           R-squared     =  0.0132
-------------+------------------------------           Adj R-squared =  0.0084
       Total |   391732982   208  1883331.64           Root MSE      =  1366.6

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         roe |   18.50119   11.12325     1.66   0.098    -3.428196    40.43057
       _cons |   963.1913   213.2403     4.52   0.000     542.7902    1383.592
------------------------------------------------------------------------------
```

****

Homework:

Show that the following three statements are true. Make sure that you clearly state all assumptions that you make and that you show all steps.

$E(\widehat{\beta_1}) = \beta_1$

$E(\widehat{\beta_0}) = \beta_0.$

$$Var(\widehat{\beta_1}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

****

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{s_x^2} = \beta_1 + (1/s_x^2)\sum_{i=1}^{n} d_i u_i, \quad \textbf{(2.52)}$$

THEOREM 2.1 (UNBIASEDNESS OF OLS)

Using Assumptions SLR.1 through SLR.4,

$$\mathrm{E}(\hat{\beta}_0) = \beta_0, \text{ and } \mathrm{E}(\hat{\beta}_1) = \beta_1 \quad \textbf{(2.53)}$$

for any values of β_0 and β_1. In other words, $\hat{\beta}_0$ is unbiased for β_0, and $\hat{\beta}_1$ is unbiased for β_1.

PROOF: In this proof, the expected values are conditional on the sample values of the independent variable. Since s_x^2 and d_i are functions only of the x_i, they are nonrandom in the conditioning. Therefore, from (2.53),

$$\mathrm{E}(\hat{\beta}_1) = \beta_1 + \mathrm{E}[(1/s_x^2)\sum_{i=1}^{n} d_i u_i] = \beta_1 + (1/s_x^2)\sum_{i=1}^{n} \mathrm{E}(d_i u_i)$$

$$= \beta_1 + (1/s_x^2)\sum_{i=1}^{n} d_i \mathrm{E}(u_i) = \beta_1 + (1/s_x^2)\sum_{i=1}^{n} d_i \cdot 0 = \beta_1,$$

where we have used the fact that the expected value of each u_i (conditional on $\{x_1, x_2, \dots, x_n\}$) is zero under Assumptions SLR.2 and SLR.3.

The proof for $\hat{\beta}_0$ is now straightforward. Average (2.48) across i to get $\bar{y} = \beta_0 + \beta_1\bar{x} + \bar{u}$, and plug this into the formula for $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \beta_0 + \beta_1\bar{x} + \bar{u} - \hat{\beta}_1\bar{x} = \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{u}.$$

Then, conditional on the values of the x_i,

$$\mathrm{E}(\hat{\beta}_0) = \beta_0 + \mathrm{E}[(\beta_1 - \hat{\beta}_1)\bar{x}] + \mathrm{E}(\bar{u}) = \beta_0 + \mathrm{E}[(\beta_1 - \hat{\beta}_1)]\bar{x},$$

since $\mathrm{E}(\bar{u}) = 0$ by Assumptions SLR.2 and SLR.3. But, we showed that $\mathrm{E}(\hat{\beta}_1) = \beta_1$, which implies that $\mathrm{E}[(\hat{\beta}_1 - \beta_1)] = 0$. Thus, $\mathrm{E}(\hat{\beta}_0) = \beta_0$. Both of these arguments are valid for any values of β_0 and β_1, and so we have established unbiasedness.

THEOREM 2.2 (SAMPLING VARIANCES OF THE OLS ESTIMATORS)
Under Assumptions SLR.1 through SLR.5,

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sigma^2/s_x^2 \quad \textbf{(2.57)}$$

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2 n^{-1}\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \textbf{(2.58)}$$

where these are conditional on the sample values $\{x_1,\ldots,x_n\}$.

PROOF: We derive the formula for $\mathrm{Var}(\hat{\beta}_1)$, leaving the other derivation as an exercise. The starting point is equation (2.52): $\hat{\beta}_1 = \beta_1 + (1/s_x^2)\sum_{i=1}^{n} d_i u_i$. Since β_1 is just a constant, and we are conditioning on the x_i, s_x^2 and $d_i = x_i - \bar{x}$ are also nonrandom. Furthermore, because the u_i are independent random variables across i (by random sampling), the variance of the sum is the sum of the variances. Using these facts, we have

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_1) &= (1/s_x^2)^2 \mathrm{Var}\left(\sum_{i=1}^{n} d_i u_i\right) = (1/s_x^2)^2\left(\sum_{i=1}^{n} d_i^2 \mathrm{Var}(u_i)\right)\\ &= (1/s_x^2)^2\left(\sum_{i=1}^{n} d_i^2\sigma^2\right) \qquad [\text{since } \mathrm{Var}(u_i) = \sigma^2 \text{ for all } i]\\ &= \sigma^2(1/s_x^2)^2\left(\sum_{i=1}^{n} d_i^2\right) = \sigma^2(1/s_x^2)^2 s_x^2 = \sigma^2/s_x^2,\end{aligned}$$

which is what we wanted to show.

Estimating the Error Variance:

We will use the following unbiased estimator of the error variance:

$$\hat{\sigma}^2 = \frac{1}{(n-2)}\sum_{i=1}^{n}\hat{u}_i^2 = \mathrm{SSR}/(n-2). \quad \textbf{(2.61)}$$

The standard error of the estimate of the slope coefficient is:

$$\mathrm{se}(\hat{\beta}_1) = \hat{\sigma}/s_x = \hat{\sigma}\Big/\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{1/2}$$