AFRE 835: Introductory Econometrics

Chapter 2: Simple Linear Regression

Spring 2017

Introduction

- In this chapter, we consider a simple regression model relating two variables (y and x).
- y denotes the *dependent variable* of interest and x is an *explanatory variable* that we believe potentially impacts y.
- We might, for example, be interested in measuring the impact of
  - Class size (x) on school performance (y);
  - Years of schooling (x) on wages earned (y);
  - Fertilizer application rates (x) on soybean yields (y);
  - Policies restricting vehicle usage (x) on local pollution levels (y);
  - Local environmental amenities (x) on housing prices (y);
  - Hospital visits (x) on health outcomes (y);
- In each of these cases, it is likely that there are many other factors influencing our variable of interest, y.
- The simple two-variable regression model helps to illustrate some basic issues in identifying the *causal* impact of x on y.

Outline

The Simple Linear Regression Model

- A common starting point is to specify a linear relationship between y and x.
- In particular, we might assume that, in the population,

$$y = \beta_0 + \beta_1 x + u \tag{1}$$

  where β_0 and β_1 denote parameters of our model and u (the *error term*) captures all other factors potentially influencing y.
- Some of factors included in u are
  1. Omitted variables (no data)
  2. Measurement errors
  3. Errors in functional form (e.g., due to the linear approximation).

The Simple Linear Regression Model (cont'd)

- This linear model says that, holding u constant, each unit change in x will change y by β_1 units; i.e.,

$$\Delta y = \beta_1 \Delta x \text{ if } \Delta u = 0 \tag{2}$$

- Written another way,

$$\frac{\Delta y}{\Delta x} = \beta_1 \text{ if } \Delta u = 0 \tag{3}$$

- The *linearity* of the model is a strong assumption, though it will often provide a good approximation on average.

The Simple Linear Regression Model (cont'd)

- The more difficult problem is estimating $\frac{\Delta y}{\Delta x} = \beta_1$ *holding everything else constant*; i.e., $\Delta u = 0$.
- Suppose, for example, that I observe data on the wages (y) and years of schooling (x) for two individuals:
  1. $\{y_1 = \$10/hour, x_1 = 6 \text{ years of schooling}\}$
  2. $\{y_2 = \$40/hour, x_2 = 16 \text{ years of schooling}\}$
- Can I infer that *all* of the difference in wages of these two individuals is due to the difference in education; i.e., that

$$\beta_1 = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{30}{10} = \$3/hour? \tag{4}$$

- To put it another way, is it likely that $\Delta u = u_2 - u_1 = 0$?

The Connection Between x and u

- In order to make *causal* inference about the relationship between changes in y and changes in x, we need to know how x and u are related.

  ...or more precisely, we need to make *assumptions* about the u's and how x and u are related (since the u's are unobservable).
- One costless assumption is the $E(u) = 0$.
  - Suppose that $E(u) = \mu_u \neq 0$.
  - Then we can always rewrite the model as

$$\begin{aligned} y &= \beta_0 + \beta_1 x + u \\ &= (\beta_0 + \mu_u) + \beta_1 x + (u - \mu_u) \\ &= \tilde{\beta}_0 + \beta_1 x + \tilde{u} \end{aligned} \tag{5}$$

  where $\tilde{\beta}_0 \equiv \beta_0 + \mu_u$ and $\tilde{u} \equiv u - \mu_u$, with $E(\tilde{u}) = E(u - \mu_u) = 0$.

Zero Conditional Mean

- We need additional assumptions on the relationship between x and u.
- One assumption would be that x and u are uncorrelated, or equivalently $Cov(x, u) = 0$.
  - This turns out not to be enough, since it does not preclude u from still being related to functions of x (e.g., x^2).
- We need the somewhat stronger assumption of *conditional mean independence*; i.e.,

$$E(u|x) = E(u). \tag{6}$$

- Combined with the zero mean assumption, this gives us the *zero conditional mean assumption*:

$$E(u|x) = 0. \tag{7}$$

- This assumption implies that u is uncorrelated with *any* function of x.

Consider Some of Our Earlier Examples

- Fertilizer application rates (x) on soybean yields (y);
- Class size (x) on school performance (y);
- Years of schooling (x) on wages earned (y);
- Policies restricting vehicle usage (x) on local pollution levels (y);
- Local environmental amenities (x) on housing prices (y);
- Hospital visits (x) on health outcomes (y);

The Population Regression Function

- Given the zero conditional mean assumption, and our linear model, we can then write that

$$
\begin{aligned}
E(y|x) &= E(\beta_0 + \beta_1 x + u|x) \\
&= \beta_0 + \beta_1 x + E(u|x) \\
&= \beta_0 + \beta_1 x. \qquad (8)
\end{aligned}
$$

- $E(y|x)$ is known as the *Population Regression Function (PRF)*.
- Given the current model, we are assuming that the *PRF* is linear in x.
- This, in turn, implies that y is, on average, a linear function of x.
- It is important to emphasize that this *does not* imply that $y = \beta_0 + \beta_1 x$.

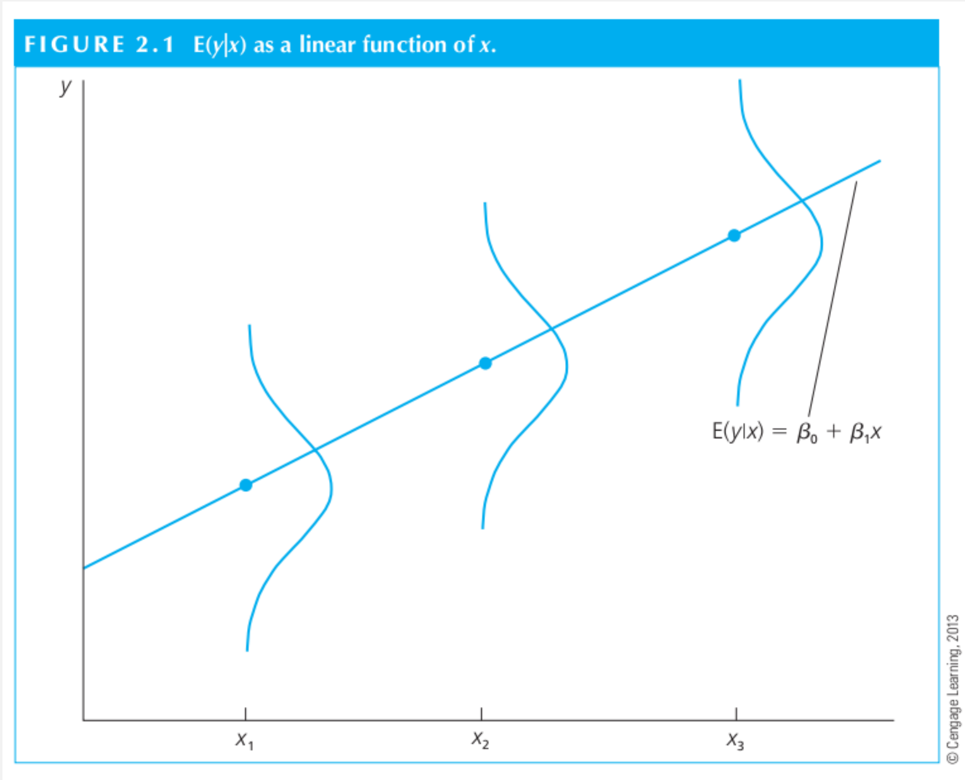
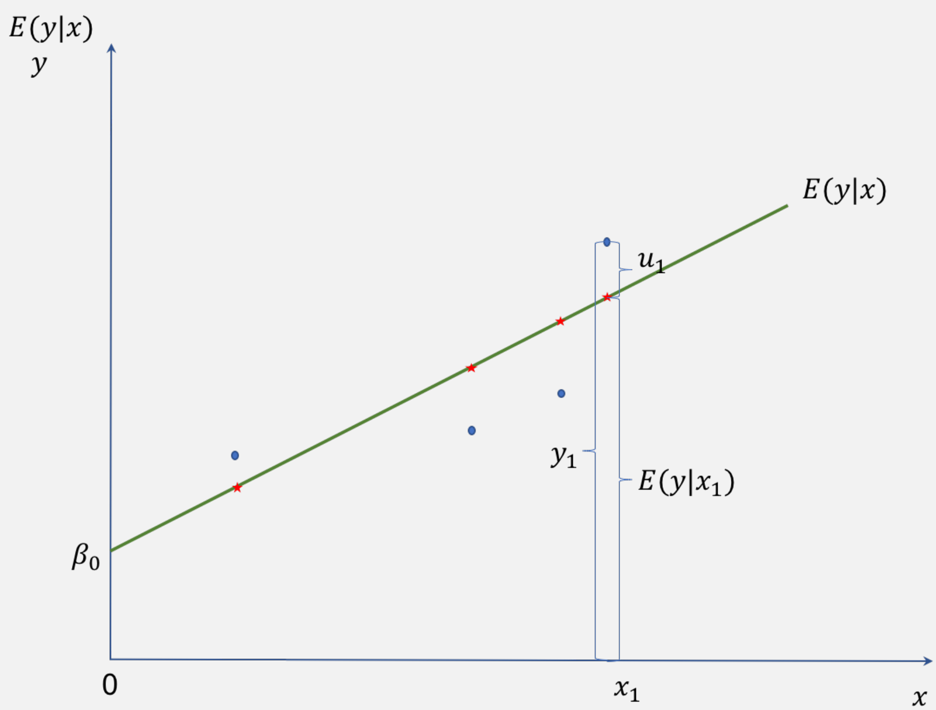


FIGURE 2.1 E(y|x) as a linear function of x.
y
E(y|x) = β0 + β1x
x1
x2
x3
© Cengage Learning, 2013


E(y|x)
y
E(y|x)
u1
y1
E(y|x1)
β0
0
x1
x

Estimation

- Given our linear PRF, we want to use a sample from the population to estimate the unknown parameters of our model; i.e., β_0 and β_1.
- An *estimator* is a rule for combining data to produce a numerical value for a population parameter; the form of the rule does not depend upon the particular sample obtained.
- An *estimate* is the numerical value taken on by an estimator for a particular sample of data.
- Important: The estimator is a random variable, whereas an estimate is not.
- Let $\{(y_i, x_i) : i = 1, \ldots, n\}$ denote a random sample from the population of interest.
- Based on our model, we know that

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{9}$$

- How do we use this information to estimate β_0 and β_1?

Choosing an Estimator

- There are many potential estimators
- For example, one might:
  - Simply graph the data and draw a "line of best fit."
  - Set $\hat{\beta}_0 = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\hat{\beta}_1 = 0$.
  - Set $\hat{\beta}_0 = 0$ and $\hat{\beta}_1 = 1$.
- We will use a variety of criterion for judging the quality of these characteristics, including:
  - unbiasedness;
  - consistency;
  - efficiency;
  - mean squared error.

The OLS Estimator

- *Ordinary Least Squares (OLS)* is a traditional estimator in the context of a linear regression model.
- To derive the OLS estimator, for a given $\hat{\beta}_0$ and $\hat{\beta}_1$, let $\hat{y}_i$ denote the *fitted value* for y when $x = x_i$, where

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \tag{10}$$

- This is our predicted value for y when $x = x_i$, since the associated error term is equal to zero on average.
- Let $\hat{u}_i$ (the *residual*) denote the difference between the true value y_i and our predicted value $\hat{y}_i$; i.e.,

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \tag{11}$$

- The OLS estimator seeks to minimize the sum of squared differences between the true value of y_i and our prediction of it $\hat{y}_i$.

The OLS Estimator (cont'd)

- Formally, the OLS estimator chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to solve

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \hat{u}_i^2 = \min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2 \tag{12}$$

- The first order conditions for this minimization problem are:

$$0 = \sum_{i=1}^{n} \left[y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right] \tag{13}$$

$$0 = \sum_{i=1}^{n} x_i \left[y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right] \tag{14}$$

- The first condition implies that

$$0 = \bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} \Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{15}$$

The OLS Estimator (cont'd)

- Substituting (15) into (14) yields

$$
\begin{aligned}
0 &= \sum_{i=1}^{n} x_i \left[y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i \right] \\
&= \sum_{i=1}^{n} x_i \left[(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}) \right] \\
&= \sum_{i=1}^{n} x_i (y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^{n} x_i (x_i - \bar{x}) \qquad (16)
\end{aligned}
$$

- Solving for $\hat{\beta}_1$ yields

$$
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad (17)
$$

- This requires $\sum_{i=1}^{n} (x_i - \bar{x})^2 \neq 0$.

An Alternative Motivation for the OLS Estimator

- An alternative approach to motivating the OLS estimator is to make use of our assumptions regarding the error term u; i.e., $E(u|x) = E(u) = 0$.
- One implication of the zero conditional mean assumption is that

$$
Cov(x, u) = E(xu) = 0 \qquad (18)
$$

- These assumptions in turn imply that:

$$
0 = E(u) = E(y - \beta_0 - \beta_1 x) \qquad (19)
$$

and

$$
0 = E(xu) = E[x(y - \beta_0 - \beta_1 x)] \qquad (20)
$$

- Note: These are assumptions regarding the underlying population.

The Method of Moments Estimator

- Since the two assumptions are expected to hold in the population, one approach to choosing our parameters is to set $\hat{\beta}_0$ and $\hat{\beta}_1$ so that the sample counterpart conditions hold.
- This is the *method of moments (MOM)* approach to estimation.
- In this case, we choose $\hat{\beta}_0$ and $\hat{\beta}_1$ such that:

$$0 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \tag{21}$$

  and

$$0 = \frac{1}{n}\sum_{i=1}^{n}\left[x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)\right] \tag{22}$$

- But these are the same as the first order conditions used to derive the OLS estimators for β_0 and β_1.
- Thus, the MOM estimator in this case is the same as the OLS estimator.

The Sample Regression Function

- Now that we have estimates for the unknown parameters we can define the *sample regression function (SRF)* as:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{23}$$

  which it the predicted value of y_i given $x = x_i$ and using the OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.
- Note: that the SRF implies that $y_i = \hat{y}_i + \hat{u}_i$.
- Using data from $WAGE1.RAW$, containing wage and education data for $n = 526$ individuals, we get

$$\widehat{wage}_i = -0.90 + 0.54 educ_i \tag{24}$$

Algebraic Properties of OLS Statistics

- There are several useful properties of OLS:
  1. The sum of the OLS residuals is zero; i.e.,
  $$\sum_{i=1}^{n} \hat{u}_i = 0. \tag{25}$$
  which is implied by the first first order condition for OLS. This *does not* imply that $\hat{u}_i = 0$ for each (or even any) i.
  2. The sample covariance between x_i and the OLS residuals ($\hat{u}_i$) is zero; i.e.,
  $$\frac{1}{n}\sum_{i=1}^{n} x_i \hat{u}_i = 0. \tag{26}$$
  which is implied by the second first order condition for OLS.
  3. The point $(\bar{x}, \bar{y})$ lies on the SRF; i.e.,
  $$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \tag{27}$$

Decomposing the Variation in y_i

- OLS can be viewed as decomposing y_i into two uncorrelated parts, with
$$y_i = \hat{y}_i + \hat{u}_i. \tag{28}$$
where
$$\frac{1}{n-1}\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = 0 \tag{29}$$
- The total sum of squares (SST) in y_i can be decomposed into the"Explained Sum of Squares" (SSE) represented by the fitted regression line $\hat{y}_i$ and the "Residual Sum of Squares" (SSR) represented by $\hat{u}_i$; i.e.,
$$SST = SSE + SSR \tag{30}$$
$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(\hat{u}_i)^2 \tag{31}$$

The Coefficient of Determination

- A common measure of how well the model "fits" the data is the *coefficient of determination*, more commonly known as the *R-squared*, where
$$R^2 = \frac{SSE}{SST} \tag{32}$$
denotes the fraction of the sample variation in y "explained" by x.
- Note: $0 \leq R^2 \leq 1$
- A value of R^2 close to 1 indicates a "good" fit, but is not necessarily an indication that the model itself is "good" or "useful."

Units of Measurement

- Changes in the units with which either the dependent or explanatory variables are measured will impact the corresponding coefficients.
- Suppose we model housing prices in dollars (*hprice*) as a function of square footage (*sqft*) and obtain the sample regression function:
$$\widehat{hprice}_i = 25.7 + 123.5 sqft_i \tag{33}$$
- This indicates that each additional square foot costs roughly \$123.50.
- If we want to measure housing prices in thousands of dollars instead (*hpricethous*), then we need to divide *hprice* by 1000, so that
$$\begin{aligned} \widehat{hpricethous}_i = \frac{\widehat{hprice}_i}{1000} &= \frac{25.7 + 123.5 sqft_i}{1000} \\ &= 0.0257 + 0.1235 sqft_i \end{aligned} \tag{34}$$
- One additional square foot increases housing prices by 0.1235 thousand dollars, which is the same \$123.50 as we got before.

Units of Measurement (cont'd)

- Changes in the units of our explanatory variable only change the parameter on that explanatory variable.
- Measuring household size in terms of hundreds of sqft (*hunsqft*), the corresponding coefficient must increase by a factor of 100.
- Specifically, we have

$$\begin{aligned}\widehat{hprice}_i &= 25.7 + 123.5 \cdot sqft_i \\ &= 25.7 + (123.5 \cdot sqft_i) \times \frac{100}{100} \\ &= 25.7 + (123.5 \times 100)\frac{sqft_i}{100} \\ &= 25.7 + 12350 \cdot hunsqft_i \qquad (35)\end{aligned}$$

- Increasing house size by 0.01 hundred square feet increases the house price by $\$12350 \cdot 0.01 = \$123.5$, the same as before.

Allowing for Nonlinearities

- When we refer to the models above as linear, the key here is that it is linear in the parameters, *not* that it is linear in y or x.
- There are a number of useful ways in which y and x can be nonlinear functions of underlying variables.
- In our model of wages as a function of education, we might want to specify $y = log(wages)$, where $log(\cdot)$ is the natural logaritm, so that

$$log(wages) = \beta_0 + \beta_1 educ + u \qquad (36)$$

- In this case, if $\Delta u = 0$, then

$$100 \cdot \beta_1 = 100 \cdot \frac{\partial log(wages)}{\partial educ} \approx \frac{\%\Delta wages}{\Delta educ} \qquad (37)$$

- Now each additional year of education has a fixed percentage impact on wages, rather than a fixed dollar impact on wages.

Allowing for Nonlinearities (cont'd)

- We can also have our explanatory variable enter in logarithmic form.
- In our model of wages as a function of education, we might want to specify $x = log(educ)$, so that

$$log(wages) = \beta_0 + \beta_1 log(educ) + u \tag{38}$$

- In this case, if $\Delta u = 0$, then

$$\beta_1 = \frac{\partial log(wages)}{\partial log(educ)} \approx \frac{\%\Delta wages}{\%\Delta educ} \tag{39}$$

- Now the slope coefficient provides a constant elasticity of wages with respect to education.

Statistical Properties of OLS Estimators

- Up to now, we have said nothing about the statistical properties of the OLS estimator.
- However, the OLS estimator is a function of data and. since the data are random drawn from the underlying population, the estimator is also random and will vary with different random samples.
- In characterizing the statistical properties, we will make use of five assumption for our simple linear regression (SLR) model.

First Four Assumptions and the Unbiasedness of OLS

SLR.1 Linear in Parameters: In the population, the dependent variable y is related to the independent variable x and the error u as

$$y = \beta_0 + \beta_1 x + u. \tag{40}$$

SLR.2 Random Sampling: We have a random sample size of n, $\{(x_i, y_i) : i = 1, \ldots, n\}$, following the population model in (40).

SLR.3 Sample Variation in the Explanatory Variable: The sample outcomes for x, namely $\{x_i : i = 1, \ldots, n\}$, are not all the same.

SLR.4 Zero Conditional Mean: The error u has an expected value of zero given any value of the explanatory variable; i.e.,

$$E(u|x) = 0 \tag{41}$$

- *Theorem 2.1*: Under assumptions SLR.1 through SLR.4, the OLS estimator is unbiased; i.e., $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$

Assumption SLR.5 and the Variances of OLS

SLR.5 Homoskedasticity: The error of u has the same variance given any value of the explanatory variable; i.e.,

$$Var(u|x) = \sigma^2 \tag{42}$$

- *Theorem 2.2* Under assumptions SLR.1 through SLR.5

$$Var(\beta_1|\boldsymbol{x}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{SST_x} \tag{43}$$

and

$$Var(\beta_0|\boldsymbol{x}) = \frac{\frac{\sigma^2}{n}\sum_{i=1}^{n}(x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\frac{\sigma^2}{n}\sum_{i=1}^{n}(x_i)^2}{SST_x} \tag{44}$$

where $SST_x \equiv \sum_{i=1}^{n}(x_i - \bar{x})^2$ and $\boldsymbol{x} = \{x_1, \ldots, x_n\}$.

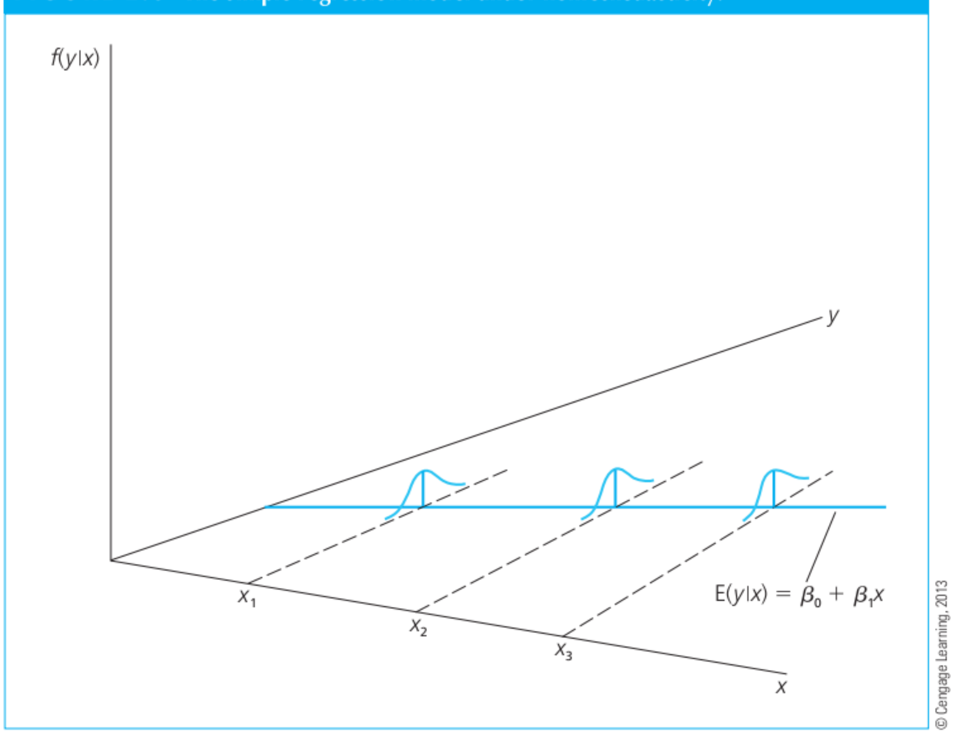
- The form of these variances makes intuitive sense.

Homoskedasticity versus Heteroskedasticity

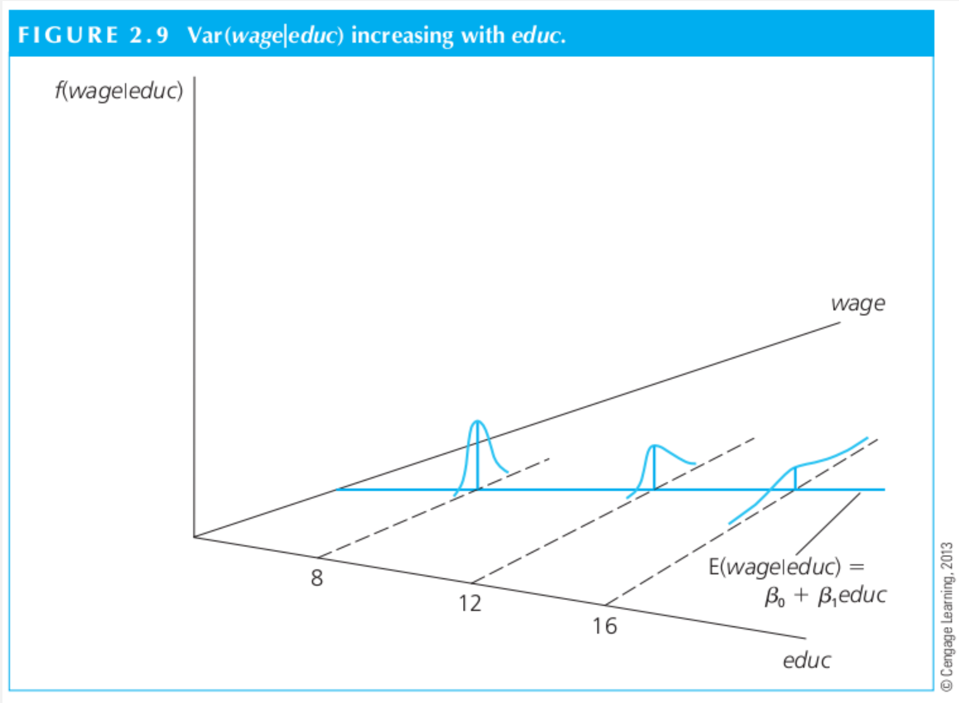
- Assumption SLR.5 is a fairly strong assumption.
- It requires that the unobserved factors impacting y have the same variability regardless of the value of what it is we do observe.
- In the case of the wage example, we are likely to observe greater variability in wage outcomes for individuals with higher education levels.
- In part, this is because the range of possible wages varies by education.
- One can obtain variances for the OLS estimator without assuming heteroskedasticity (referred to as *robust variances*), but the formula's are more complicated.

Homoskedasticity

FIGURE 2.8 The simple regression model under homoskedasticity.



Heteroskedasticity



Estimating the Error Variance

- The formulas for the OLS estimator variances are useful in terms of understanding how these variances are impacted by various factors, including
  - Sample size;
  - Variability in the unknown error terms; and
  - Variability in the explanatory variable.
- But in practice, we rarely know σ^2 and must come up with an estimator for it.
- *If* we observed the error terms, u_i's, the task would be easy, since

$$E(u) = \sigma^2. \tag{45}$$

- This suggests an unbiased estimator for σ^2 of

$$\frac{1}{n}\sum_{i=1}^{n} u_i^2 \tag{46}$$

Estimating the Error Variance (cont'd)

- Without the errors themselves, the residuals from our OLS regression will be helpful.
- Note that the errors and residuals are *not* the same thing:

$$\begin{aligned} \hat{u}_i &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \\ &= u_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1) x_i \end{aligned} \tag{47}$$

- Since the OLS estimator is unbiased, the difference between the errors and the residuals is zero.
- An unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{u}_i^2 = \frac{SSR}{n-2} \tag{48}$$

- *Theorem 2.3*: Under assumptions SLR.1 through SLR.5, $\mathrm{E}(\hat{\sigma}^2) = \sigma^2$; i.e., $\hat{\sigma}^2$ is an unbiased estimator of σ^2.

The Standard Error of the Regression (SER)

- We are often interested in constructing the standard deviation of our parameter estimates.
- For example, in the case of the slope parameter

$$sd(\hat{\beta}_1) = \frac{\sigma}{SST_x} \tag{49}$$

- An estimator of $sd(\hat{\beta}_1)$ replaces σ with $\hat{\sigma} \equiv \sqrt{\hat{\sigma}^2}$ (the so-called *standard error of the regression, or SER).*
- The SER is not an unbiased estimator of σ, but it is consistent.
- The resulting estimator of $sd(\hat{\beta}_1)$ is

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{SST_x} \tag{50}$$

  the standard error of $\hat{\beta}_1$.