Topic 3: The Bivariate Linear Regression Model

- In our review of probability and statistics in Topic 2 we saw that given the joint probability distribution of two random variables, X and Y, we can derive:
  - The marginal distribution of X.
  - The marginal distribution of Y.
  - The conditional distribution of Y (the distribution Y conditional on X assuming a particular value).
- We also saw that once we have derived the conditional distribution of Y, we can derive the conditional mean of Y. That is, the mean of Y given that X has assumed a particular value.
- For example, the average number of bathrooms in the population of houses with 3 bathrooms.
- In a **bivariate regression model** we have two random variables, the dependent variable, Y, and a single explanatory variable, X.

- The relationship which is of greatest interest to us is that between the average value of Y and the value of the explanatory variable, X. That is, $E(Y|X = x)$.
- For example, a labor economist may be interested in estimating the average wage of the population of individuals who have 12 years of education.
- In this example, the conditional mean of interest would be $E(wage|education = 12)$.
- Our hypothetical labor economist may also be interested in estimating how the average wage changes in response to changes in years of education.
- For example, she may wish to estimate

$$E(wage|education = 12) - E(wage|education = 11).$$

- Given that the quantity of interest is $E(Y|X = x)$, we generally start our empirical analysis by choosing a mathematical model for $E(Y|X = x)$.

Note: Strictly speaking, we should use the notation $E(Y|X = x)$ to denote the conditional mean of Y.However, we will follow the common convention in econometrics and use the more compact notation $E(y|x)$.

- In the **bivariate linear regression model** (BLRM) we assume that $E(y_i|x_i)$ is a linear function of both the parameters and the explanatory variable X. That is,

$$E(y_i|x_i) = \beta_0 + \beta_1 x_i, i = 1, 2, ..., n. \quad (1)$$

- The key property of (1) is that $E(y_i|x_i)$ is **linear in the unknown parameters** β_0 and β_1.

- Were we to relax the assumptions that $E(y_i|x_i)$ is linear in x_i and assume, for example, that

$$E(y_i|x_i) = \beta_0 + \beta_1 x_i^2, i = 1, 2, ..., n, \quad (2)$$

  the model would still be a linear regression model.
- However, as we will see later in the unit, the interpretation of β_1 in (2) is different from the interpretation of β_1 in (1).
- The model in (1) assumes that there is an exact linear relationship between $E(y_i|x_i)$ and the value assumed by X.
- The assumption that $E(y_i|x_i)$ is linear in both the parameters and x_i is illustrated in Figure 1 below.

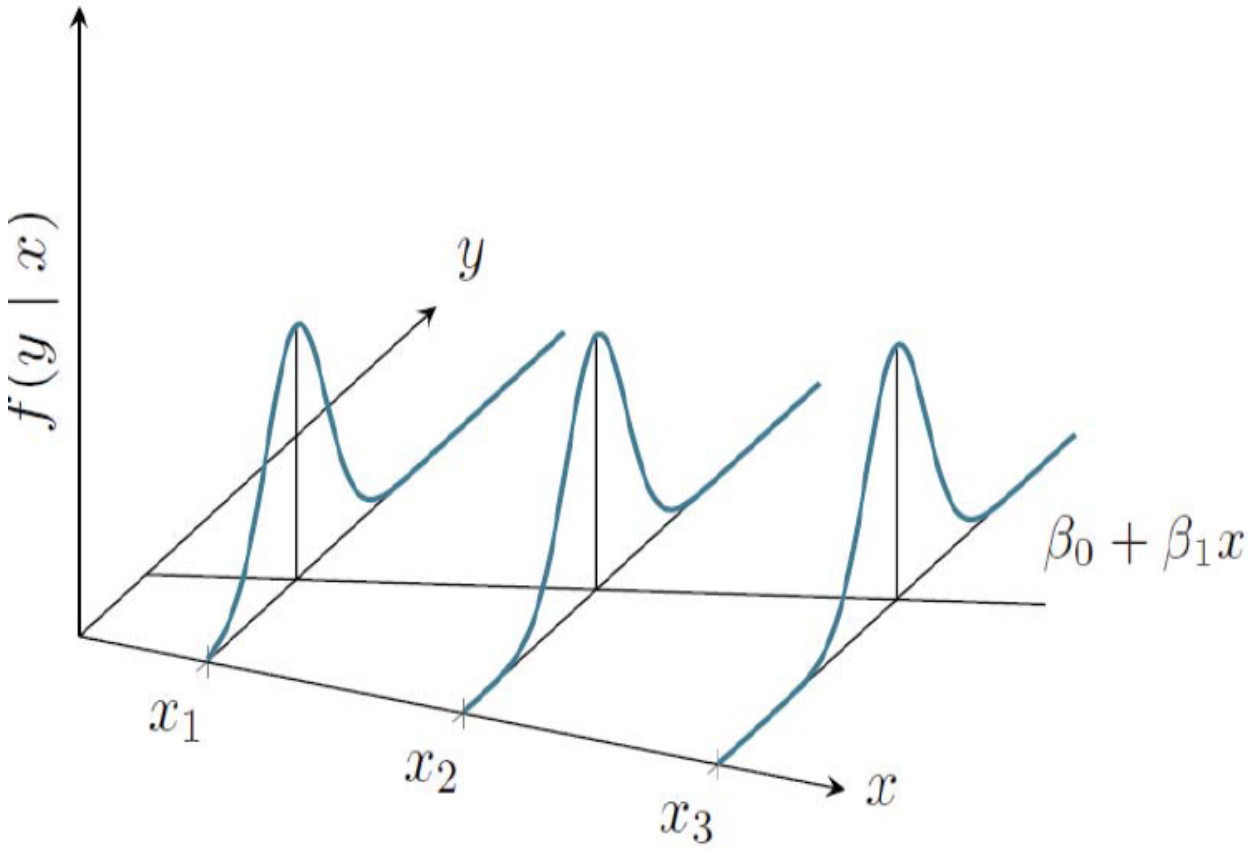

Figure: 1

- From (1) notice that

$$\begin{aligned} E(y_i|x_i+1) - E(y_i|x_i) &= \beta_0 + \beta_1(x_i+1) - (\beta_0 + \beta_1 x_i) \\ &= \beta_0 + \beta_1 x_i + \beta_1 - \beta_0 - \beta_1 x_i \\ &= \beta_1. \end{aligned}$$

- Since

$$E(y_i|x_i+1) - E(y_i|x_i) = \beta_1, \tag{3}$$

it follows that in the BLRM

$$E(y_i|x_i) = \beta_0 + \beta_1 x_i, i = 1, 2, ..., n, \tag{1}$$

the coefficient β_1 measures the change in the conditional mean of Y in response to a one unit change in X.

- For example, if Y denotes individual wages and X denotes years of education, β_1 measures the difference between the average wage of the population of individuals with x years of education and the average wage of the population with $x+1$ years of education.
- For those of you who are familiar with elementary calculus,

$$\frac{dE(y_i|x_i)}{dx_i} = \beta_1. \tag{4}$$

- Of course, it is highly unlikely that a randomly chosen value of Y, say y_i, would be exactly equal to $E(y_i|x_i)$.

- Let u_i denote the deviation of y_i from $E(y_i|x_i)$. Then we may write

$$u_i = y_i - E(y_i|x_i), \tag{5}$$

or rearranging (5)

$$y_i = E(y_i|x_i) + u_i \tag{6}$$

- Taking conditional expectations on both sides of (6) we obtain

$$\begin{aligned} E(y_i|x_i) &= E[E(y_i|x_i) + u_i] \\ &= E(y_i|x_i) + E(u_i|x_i), \end{aligned}$$

which implies that

$$E(u_i|x_i) = E(y_i|x_i) - E(y_i|x_i) = 0. \tag{7}$$

- Finally, substituting

$$E(y_i|x_i) = \beta_0 + \beta_1 x_i \tag{1}$$

into

$$y_i = E(y_i|x_i) + u_i \tag{6}$$

we obtain

$$y_i = \beta_0 + \beta_1 x_i + u_i . i = 1, 2, ...n, \tag{8}$$

where

$$E(u_i|x_i) = 0.$$

- Equation (8) is the standard representation of the bivariate linear regression model.
- From (8)

$$\frac{dy_i}{dx_i} = \beta_1, i = 1, 2, ...n. \tag{9}$$

- Equation (9) states that the parameter β_1 measures the change in y_i arising from a small change in x_i.
- As we mentioned in Topic 1,econometricians call β_1 the **partial effect** or **marginal effect** of x_i on y_i.
- Note from (9) that

$$\frac{dy_1}{dx_1} = \frac{dy_2}{dx_2} = ... = \frac{dy_n}{dx_n} = \beta_1,$$

implying that the marginal effect of X on Y is constant across observations.

- Assume that we have n observations on x and y and we wish to use these n observations to estimate β_0 and β_1 in the regression equation

$$y_i = \beta_0 + \beta_1 x_i + u_i . i = 1, 2, ...n. \tag{10}$$

- Of course, we wish to obtain estimates of β_0 and $\beta_1$which are in some sense "accurate".
- Recall that in the bivariate linear regression model

$$E(y_i|x_i) = \beta_0 + \beta_1 x_i. \tag{5}$$

- Let

$$\widehat{y}_i = b_0 + b_1 x_i, \tag{11}$$

where b_0 is an estimate of β_0 and b_1 is an estimate of β_1.

- That is, $\widehat{y}_i$ is the estimated or predicted value of y_i when we use b_0 and b_1 in place of β_0 and β_1.
- Let

$$\begin{aligned} \widehat{u}_i &= y_i - \widehat{y}_i \\ &= y_i - b_0 - b_1 x_i \text{ (using (10))}. \end{aligned} \tag{12}$$

- That is, $\widehat{u}_i$ is the **prediction error** or **residual** we obtain when we use $\widehat{y}_i$ as our prediction of y_i.

- The OLS estimator is derived by choosing as our estimators of β_0 and β_1 those values of b_0 and b_1 which minimize

$$\sum_{i=1}^{n} \widehat{u}_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2. \tag{13}$$

  That is, we choose b_0 and b_1 to minimize the **sum of the squared residuals** (or equivalently, the sum of the squared prediction errors) given by (13).
- It seems intuitively plausible that choosing b_0 and b_1 to minimize the sum of the squared residuals in (13) should provide estimates of β_0 and β_1 that are "accurate".
- We will see later in the unit that, under certain assumptions, this intuition turns out to be correct

- **Question:** Why not choose b_0 and b_1 to minimize $\sum_{i=1}^{n} \widehat{u}_i$ rather than to minimize $\sum_{i=1}^{n} \widehat{u}_i^2$?
- The problem of choosing b_0 and b_1 to minimize

$$SSR(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \tag{14}$$

is a standard calculus problem. (The notion $SSR(b_0, b_1)$ in (14) is used to emphasize that the value of SSR depends on the values of b_0 and b_1).

- The values of b_0 and b_1 which minimize $SSR(b_0, b_1)$ are found by setting the first-order partial derivatives of $SSR(b_0, b_1)$ with respect to b_0 and b_1 equal to zero and solving the resulting equations for b_0 and b_1.

- Partially differentiating (14) with respect to b_0 and b_1 and setting the partial derivatives equal to zero we obtain

$$\left.\frac{\partial SSR(b_0, b_1)}{\partial b_0}\right|_{\widehat{\beta}_0, \widehat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0 \qquad (15)$$

$$\left.\frac{\partial SSR(b_0, b_1)}{\partial b_1}\right|_{\widehat{\beta}_0, \widehat{\beta}_1} = -2\sum_{i=1}^{n} x_i(y_i - b_0 - b_1 x_i) = 0 \qquad (16)$$

- Let $\widehat{\beta}_0$ and $\widehat{\beta}_1$ denote the values of b_0 and b_1 which satisfy (15) and (16).

- It is straightforward to show that

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \tag{17}$$

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}, \tag{18}$$

where

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

(See Wooldridge for the algebraic details).

- Equations (17) and (18) are respectively the formulas for the **OLS estimators** of β_1 and β_0 in the bivariate linear regression model

$$y_i = \beta_0 + \beta_1 x_i + u_i . i = 1, 2, ...n. \tag{10}$$

- Equations (17) and (18) provide us with formulas for the OLS estimators of β_1 and β_0 respectively, which depend only on the sample values of X and Y.
- Consequently, one we have collected our sample of observations on x and y, we can plug those values into (17) and (18) and obtain **estimates** of β_1 and β_0 respectively.
- Recall from Topic 1 the distinction between an estimator and an estimate of a parameter of interest.
- **Estimator:**

- **An estimator** of a unknown parameter **is a formula** which specifies how to use the sample observations to obtain an estimate of the parameter.
- Because X and Y are random variables, their values are unknown **before** we collect our sample. Since the estimator depends on the values assumed by X and Y, this makes t**he estimator a random variable** also.
- Therefore, before we collect our sample $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are both random variables whose values are unknown.

- **Estimate:**
  - **An estimate** of an unknown parameter **is a number** we get when we collect our sample and plug the samples values into the formula for the estimator.
  - We cannot compute an estimate until after we have collected our sample.

- Since the sample values of x and y are known at the time we compute our estimate, **the estimate in not a random variable**.

- The variables

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i, i = 1, 2, ..., n$$

are called the **predicted values** or **fitted values** of y based on the OLS estimators of β_1 and β_0.

- The estimated regression coefficients are interpreted as follows:
  - $\widehat{\beta}_0$ is the predicted value of y_i when

$$x_i = 0.$$

Whether or not the case in which

$$x_i = 0$$

has any practical relevance depends on the application.

- For example, if x_i denotes the IQ score of individual i, then

$$x_i = 0$$

  does not have any practical relevance.
- As we have seen, in the BLRM

$$y_i = \beta_0 + \beta_1 x_i + u_i . i = 1, 2, ...n, \tag{10}$$

  β_1 measures the change in y_i in response to a small change in x_i.
- Therefore, $\widehat{\beta}_1$ is the predicted (or estimated) change in y_i in response to a small change in x_i.

- The variables

$$\begin{aligned} \widehat{u}_i &= y_i - \widehat{y}_i, i = 1, 2, ..., n \\ &= y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i, i = 1, 2, ..., n \end{aligned}$$

are called the **OLS residuals**, since they are the prediction errors or residuals we obtain when we use the OLS estimates to predict y_i.

- Note that

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} \\ &= \frac{\frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2} \\ &= \frac{\widehat{\sigma}_{xy}}{\widehat{\sigma}_x^2}, \end{aligned} \tag{19}$$

where

$$\widehat{\sigma}_{xy} = \text{sample covariance between } X \text{ and } Y$$

and

$$\widehat{\sigma}_x^2 = \text{sample variance of } X.$$

- Equation (19) states that in the bivariate linear regression model, the OLS estimator of β_1 is equal to the sample covariance between X and Y divided by the sample variance of X.
- Rearranging

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}. \tag{18}$$

we obtain

$$\overline{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{x}. \tag{20}$$

- As we shall see when we study the statistical properties of the OLS estimator, the intuition that choosing estimators which minimize the SSR should produce estimators that are in some sense accurate, turns out to be correct under certain assumptions.
- The intuition behind the OLS estimator is illustrated in Figure 2 below.

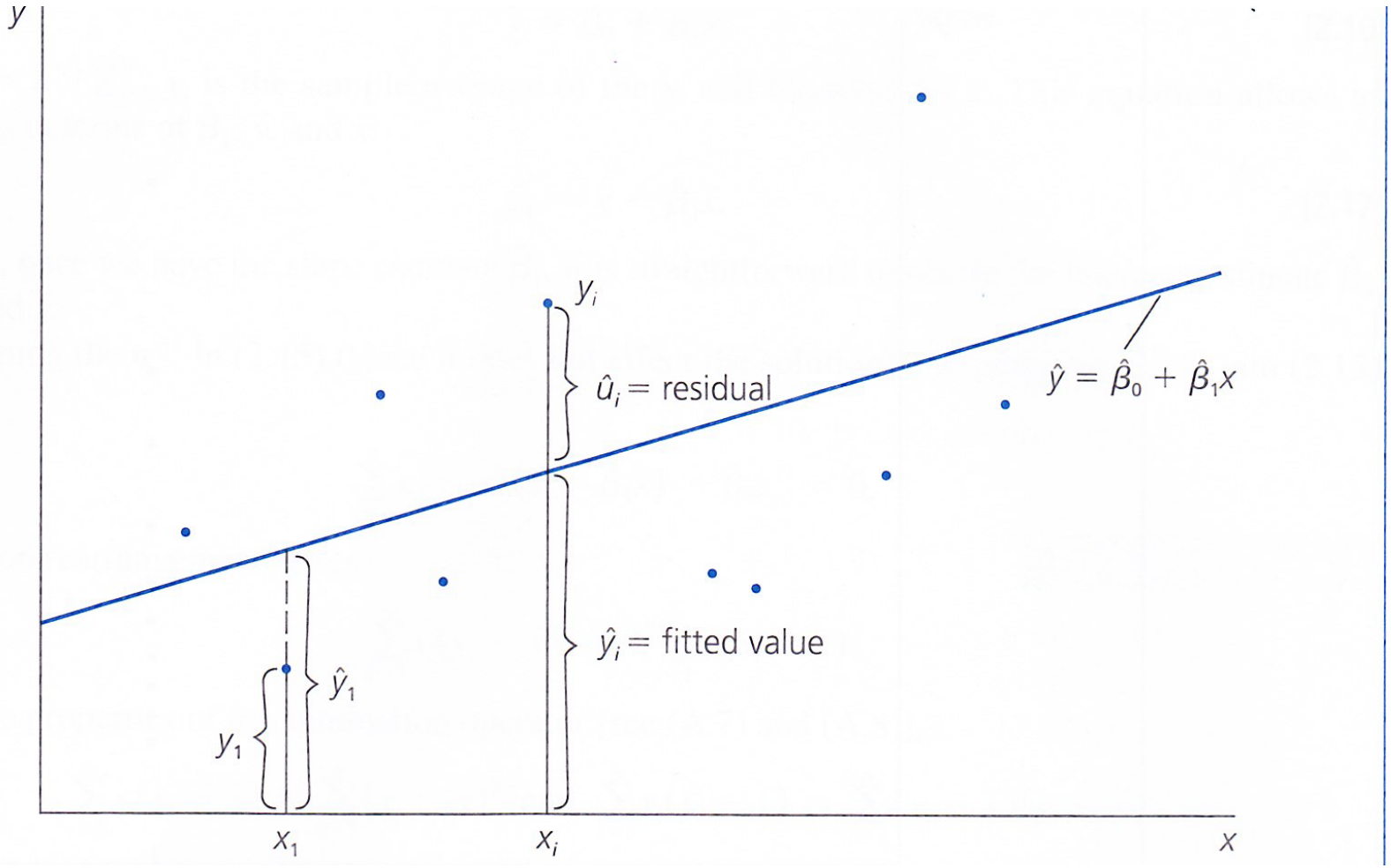

Figure: 2

The OLS estimator chooses the straight line which "best fits the data" in

- Suppose that we wish to investigate the relationship between education and wages.
- We select a random sample of 935 individuals and collect information on their weekly wage measured in dollars and their education measures in years.
- What kind of a data set do we have?
- As a starting point we estimate the following linear regression equation by OLS:

$$wage_i = \beta_0 + \beta_1 educ_i + u_i, i = 1, 2, ...935. \quad (21)$$

- When (21) is estimated in Eviews by OLS we obtain the results reported in Figure 3 below.

Dependent Variable: WAGE
Method: Least Squares
Sample: 1 935
Included observations: 935

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	146.9524	77.71496	1.890916	0.0589
EDUC	60.21428	5.694982	10.57322	0.0000

R-squared	0.107000	Mean dependent var	957.9455
Adjusted R-squared	0.106043	S.D. dependent var	404.3608
S.E. of regression	382.3203	Akaike info criterion	14.73253
Sum squared resid	1.36E+08	Schwarz criterion	14.74289

Figure: 3

- Based on the output reported in Figure 3 we can deduce the following conclusions:

- An individual's predicted weekly wage, given their education in years, is given by

$$\begin{aligned} \widehat{wage}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 educ_i, i = 1, 2, ..., n \\ &= 146.95 + 60.21 educ_i, i = 1, 2, ...n, \end{aligned}$$

- The predicted weekly wage of an individual with no education is

$$\begin{aligned} \widehat{wage}_i &= 146.95 + 60.21(0) \\ &= \$146.95. \end{aligned}$$

- The predicted wage of an individual with 10 years of education is

$$\begin{aligned} \widehat{wage}_i &= 146.95 + 60.21(10) \\ &= \$749.05. \end{aligned}$$

- Since
$$\widehat{\beta}_1 = 60.21,$$
we predict that, on average, an extra year of education will increase an individual's wage by \$60.21.
- That is, we predict that the average weekly wage of the population of individuals with x years of education will exceed the average weekly wage of the population of individuals with x-1 years of education by \$60.21.
- We predict that, on average, an extra two years of education will increase an individual's wage by
$$2\widehat{\beta}_1 = 60.21.x2 = \$120.42.$$
- That is, we predict that the average weekly wage of the population of individuals with x years of education will exceed the average weekly wage of the population of individuals with x-2 years of education by \$120.42.

- Can we interpret $\widehat{\beta}_1$ as the estimated **causal effect** of education on average wages?
- That is, can we conclude that an extra year of education is estimated to **cause** an individual's weekly wage to increase by \$60.21?
- To answer this question, recall the properties of a randomized controlled experiment discussed in Topic 1.
- In the present example, the treatment is years of education, and different individuals in our sample **have chosen** to receive different levels of the treatment.
- We must ask ourselves whether or not there are systematic differences between the individuals who have chosen different levels of the treatment (years of education) and if there are, do these differences wholly or partially explain the differences in wages of different individuals?

- What do you think?

- In this section we state without proof some **algebraic properties** of OLS estimation which **always hold** when we estimate the linear regression equation

$$y_i = \beta_0 + \beta_1 x_i + u_i . i = 1, 2, ...n, \tag{10}$$

  by the method of OLS.
- These properties are algebraic properties in the sense that they follow from the algebra OLS.
- In particular, these algebraic properties do not depend on the data or how the data has been collected. They hold for any data set.

P1 **The OLS residuals sum to zero.** That is,

$$\sum_{i=1}^{n} \widehat{u}_i = 0, \tag{22}$$

where

$$\begin{aligned} \widehat{u}_i &= y_i - \widehat{y}_i, i = 1, 2, ..., n \\ &= y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i, i = 1, 2, ..., n, \end{aligned}$$

and $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the OLS estimates of β_0 and β_1 respectively.

- Note that the fact that

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

immediately implies that

$$\overline{\widehat{u}} = \frac{1}{n} \sum_{i=1}^{n} \widehat{u}_i = 0. \tag{23}$$

Equation (23) states that the sample mean of the OLS residuals is equal to zero.

P2 The OLS residual are orthogonal to x. That is,

$$\sum_{i=1}^{n} x_i \widehat{u}_i = 0. \tag{24}$$

- If we define x' to be the (1xn) row vector

$$x' = (x_1, x_2, ..., x_n)$$

and we define $\widehat{u}$ to be the (nx1) column vector

$$\widehat{u} = \begin{pmatrix} \widehat{u}_1 \\ \widehat{u}_2 \\ . \\ . \\ \widehat{u}_n \end{pmatrix},$$

then

$$x'\widehat{u} = (x_1, x_2, ..., x_n) \begin{pmatrix} \widehat{u}_1 \\ \widehat{u}_2 \\ . \\ . \\ \widehat{u}_n \end{pmatrix} = \sum_{i=1}^{n} x_i \widehat{u}_i = 0.$$

Therefore,

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

may be written more compactly as

$$x' \widehat{u} = 0. \tag{25}$$

- Whenever two (nx1) vectors, v and w, have the property that

$$v' w = 0,$$

we say that the vectors v and w are **orthogonal** to each other.

- Therefore, another way of stating that

$$x' \widehat{u} = 0 \tag{25}$$

is to say that, **the vector of OLS residuals is orthogonal to the vector x.**

P3 In the bivariate linear regression model

$$\bar{y} = \bar{\widehat{y}} \tag{26}$$

where

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,\ \bar{\widehat{y}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{y}_i. \tag{27}$$

- P3 states that when we estimate

$$y_i = \beta_0 + \beta_1 x_i + u_i . i = 1, 2, ...n, \tag{10}$$

by OLS, the sample mean of y and the sample mean of $\widehat{y}$ are equal.

- In the bivariate linear regression model, we can decompose the total variation in the dependent variable **in our sample** into a component that is explained, or predicted, by the variation in the explanatory variable, x, and a component which is not explained, or predicted, by x.
- We use the notation R^2 to denote the proportion of the **sample variation** in y that is explained, or predicted, by x.
- Note that if there was no variation in y in our sample then

$$y_1 = y_2 = ... = y_n \Rightarrow y_i = \overline{y}, i = 1, 2, ...n.$$

In this case

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = 0.$$

- Therefore, the quantity $\sum_{i=1}^{n}(y_i - \overline{y})^2$ may be used as a measure of the total variation in y in our sample.
- Using the same logic, the quantity $\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$ may be used as a measure of the variation in y predicted by our estimated model.
- Using P1, P2 and P3 above, it can be shown that

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}\widehat{u}_i^2. \tag{28}$$

- Define

$$\begin{aligned} \text{total sum of squares} &= SST = \sum_{i=1}^{n}(y_i - \overline{y})^2, \\ \text{explained sum of squares} &= SSE = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2, \\ \text{residual sum of squares} &= SSR = \sum_{i=1}^{n}\widehat{u}_i^2. \end{aligned}$$

- Then

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}\widehat{u}_i^2 \tag{28}$$

may be written as

$$SST = SSE + SSR. \tag{29}$$

- Equation (29) states that the total variation in the dependent variable, SST, may be decomposed (broken down) into a component, SSE, which is explained or predicted by our estimated model, and a component SSR which is not explained or predicted by our estimated model.
- In words, (29) states that "the total sum of squares is equal to the explained sum of squares plus the residual sum of squares".
- Dividing both sides of (29) by SST we obtain

$$1 = \frac{SSE}{SST} + \frac{SSR}{SST},$$

or

$$\frac{SSE}{SST} = 1 - \frac{SSR}{SST},$$

or

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

where

$$R^2 = \frac{SSE}{SST}. \tag{30}$$

- R^2 is called the **coefficient of determination** and, as long as we include an intercept, $\beta_{0,}$ in the regression equation

$$y_i = \beta_0 + \beta_1 x_i + u_i . i = 1, 2, ...n, \tag{10}$$

it can be shown that

$$0 \leq R^2 \leq 1.$$

- R^2 is a measure of the proportion of the variation in the dependent variable in our sample (**the sample variation**) that is "explained or predicted" by the explanatory variable x.
- The closer R^2 is to 1, the greater the proportion of the sample variation in the dependent variable that is explained by x.
- R^2 is sometimes referred to as a measure of **goodness of fit**, since it is a measure of "how close" the predicted values of y are to the actual values of y.
- As we have seen above, when we estimate

$$wage_i = \beta_0 + \beta_1 educ_i + u_i, i = 1, 2, ...935. \quad (31)$$

by OLS in Eviews we obtain the output reported in Figure 3 below

Dependent Variable: WAGE
Method: Least Squares
Sample: 1 935
Included observations: 935

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	146.9524	77.71496	1.890916	0.0589
EDUC	60.21428	5.694982	10.57322	0.0000

R-squared	0.107000	Mean dependent var	957.9455
Adjusted R-squared	0.106043	S.D. dependent var	404.3608
S.E. of regression	382.3203	Akaike info criterion	14.73253
Sum squared resid	1.36E+08	Schwarz criterion	14.74289

Figure: 3

- Notice that

$$R^2 = 0.107.$$

  Therefore, the regressor education explains 10.7% of the sample variation in wages, implying that approximately 89% of the sample variation in wages is explained by other factors.
- A reasonable question to ask is "does a low R^2 mean that our linear regression model is useless"?
- The answer is that it depends on the purpose for which the model is being estimated;
  - If the purpose of the model is to provide a comprehensive explanation of the sample variation in y, then a low R^2 is a problem.
  - However if, as is often the case, the purpose of the model is to obtain an accurate estimate of the effect of x on y, then a low R^2 need not be a concern.