Topic 8

1. Regression with Logs
2. Quadratic Regression
3. Model Selection Criteria

- Up to this point we have confined our discussion of regression models to the linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i, \ \ i = 1, 2, ...n. \quad (1)$$

  This model is linear in the dependent variable, each of the regressors and the regression coefficients.

- There are several very common and useful generalizations of (1) which relax the assumption that the model is linear in the dependent variable and/or the regressors, but retain the assumption that the model is linear in the regression coefficients.
- As long as a regression model is linear in the regression coefficients, it can be estimated by OLS and the hypothesis testing procedures which we discussed in previous lectures remain valid when A1-A5 hold.

- Let z be a variable which can only assume positive values.
- Let $\log(z)$ denote the natural logarithm of z and assume that the value of z changes from z_0 to z_1.
- Then

$$
\begin{aligned}
\Delta \log(z) &= \log(z_1) - \log(z_0) \\
&= \log\left(\frac{z_1}{z_0}\right) \\
&= \log\left(1 - 1 + \frac{z_1}{z_0}\right) \\
&= \log\left(1 + \frac{z_1}{z_0} - \frac{z_0}{z_0}\right) \\
&= \log\left(1 + \frac{z_1 - z_0}{z_0}\right). \qquad (2)
\end{aligned}
$$

- Now, for c "small", it can be shown that

$$\log(1+c) \approx c. \tag{3}$$

- For example, when

$$c = 0.1,$$

$$\begin{aligned} \log(1+c) &= \log(1+0.1) \\ &= \log(1.1) \\ &= 0.095 \\ &\approx 0.1. \end{aligned}$$

- Note that this approximation is only accurate for "small" values of c. For example, when

$$c = 3,$$

$$\begin{aligned} \log(1+c) &= \log(1+3) \\ &= \log(4) \\ &= 1.39, \end{aligned}$$

which is not approximately equal to 3.

- Recall that

$$\Delta \log(z) = \log\left(1 + \frac{z_1 - z_0}{z_0}\right). \tag{2}$$

Using the fact that, for small values of c,

$$\log(1 + c) \approx c, \tag{3}$$

it follows that, for small changes in z,

$$\begin{aligned}\Delta \log(z) &= \log\left(1 + \frac{z_1 - z_0}{z_0}\right) \\ &\approx \frac{z_1 - z_0}{z_0}.\end{aligned}$$

That is, for small changes in z,

$$\Delta \log(z) \approx \frac{z_1 - z_0}{z_0}. \tag{4}$$

- It immediately follows from (4) that

$$100\Delta \log(z) \approx 100\left(\frac{z_1 - z_0}{z_0}\right) = \%\Delta z.$$

- In summary, for "small" changes in the value of the variable z

$$100\Delta \log(z) \approx \%\Delta z. \tag{5}$$

- For example, when z increases from 1 to 1.1

$$\%\Delta z = 100\left[\frac{1.1 - 1}{1}\right] = 10.$$

The approximate change in z

$$\begin{aligned} 100\Delta \log(z) &= 100[\log(1.1) - \log(1)] \\ &= 100(0.0953 - 0) \\ &= 9.53. \end{aligned}$$

- The first generalization of the linear regression model which we consider is called the **log-level model** and is given by

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \tag{6}$$

  where y is a variable which can assume only positive values and where, to economize on notation, we have assumed that $k = 2$.
- The model given by (6) is called the "log-level model" because the dependent variable is measured in logs and the regressors are measured in levels.
- Even though the dependent variable in (6) is the log of y rather than the level of y, we can estimate (6) by OLS because it is linear in the regression coefficients.

- It follows from

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \tag{6}$$

that when we have a small change in x_1 holding x_2 fixed, the resulting change in $\log(y)$ is given by

$$\Delta \log(y) = \beta_1 \Delta x_1,$$

which implies that

$$\frac{\Delta \log(y)}{\Delta x_1} = \beta_1.$$

Therefore,

$$\frac{100\Delta \log(y)}{\Delta x_1} = 100\beta_1,$$

or, using the result that

$$100\Delta\log(y) \approx \%\Delta y, \tag{7}$$

we obtain

$$\frac{\%\Delta y}{\Delta x_1} = 100\beta_1. \tag{8}$$

- Therefore, when we have a one unit change in x_1 holding x_2 fixed, (8) implies that

$$\%\Delta y \approx 100\beta_1.$$

- Similarly, when we have a one unit change in x_2 holding x_1 fixed, (8) implies that

$$\%\Delta y \approx 100\beta_2.$$

- Note carefully that even though the dependent variable in

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \tag{6}$$

is $\log(y)$, the coefficient β_1 measures the % change in the **level** of y (not in $\log(y)$) in response to a one unit change in the level of x_1, holding x_2 constant.

- Consider the log-level model

$$Log(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 \exp er_i + \beta_3 IQ_i + u_i, i = 1, 2, \ldots 935. \tag{9}$$

- When we estimate (9) by OLS we obtain the output reported in Figure 1 below.

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

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5.198085	0.121543	42.76759	0.0000
EDUC	0.057108	0.007348	7.771960	0.0000
EXPER	0.019525	0.003244	6.018132	0.0000
IQ	0.005786	0.000980	5.905770	0.0000

R-squared	0.162244	Mean dependent var	6.779004
Adjusted R-squared	0.159545	S.D. dependent var	0.421144
S.E. of regression	0.386089	Akaike info criterion	0.938773
Sum squared resid	138.7795	Schwarz criterion	0.959481
Log likelihood	-434.8764	Hannan-Quinn criter.	0.946669
F-statistic	60.10079	Durbin-Watson stat	1.811646
Prob(F-statistic)	0.000000		

Figure: 1

- Based on the output in Figure 1, we conclude that:

- Controlling for experience and IQ score, an extra year of education is predicted to increase the average wage by approximately

$$100(0.057) = 5.7\%.$$

- For example, we estimate (predict) that the average wage of the population of individuals with a given level of experience and a given IQ score who have 10 years of education, will be approximately 5.7% higher than the average wage of the population of individuals with the same level of experience and IQ score who have 9 years of education.
- Controlling for education and IQ score, we estimate (predict) that an extra year of experience increases the average wage by approximately

$$100(0.0196) = 1.96\%.$$

- Controlling for education and experience, we estimate (predict) that a one unit increase in IQ score increases the average wage by approximately
$$100(0.006) = 0.6\%.$$
- For example, we estimate (predict) that the average wage of the population of individuals with a given level of education and experience and an IQ score 100 will be approximately 0.6% higher than the average wage of the population of individuals with the same level of education and experience and an IQ score of 99.

- In many applications, the percentage change in the dependent variable in response to a one unit change in a regressor is easier to interpret than is the change in the level of the dependent variable.

2 Regression with Logs VIII

2.2 The log-level model

- In addition, a model specification which allows for a constant percentage change in the dependent variable in response to a one unit change in a regressor may be more plausible than a model which implies a constant change in the level of the dependent variable.
- For example, a log-level specification for wages, which predicts a constant percentage change in wages in response to an extra year of education, is more plausible than a level-level specification, which predicts a constant change in the level of wages in response to an extra year of education.
- In summary, the log-level model is an attractive specification in situations in which it makes more sense to estimate the % change in the value of the dependent variable rather than the change in the level of the dependent variable, in response to a one unit change in the value of a regressor.

- Another popular variant of the linear regression model is the **level-log** model.
- The definitive characteristic of the level-log model is that the dependent variable is measured in levels and one or more of the regressors is measured in logs.
- The regression equation given by

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + u \tag{10}$$

  is an example of a level-log model, because the dependent variable is measured in levels and the regressor x_1 is measured in logs.
- Even though we have the log of x_1 rather than the level of x_1 on the right-hand side of (10), we can estimate (10) by OLS because it is linear in the regression coefficients.

- Holding x_2 constant, it follows from (10) that

$$\Delta y = \beta_1 \Delta \log(x_1)$$

$\Rightarrow$

$$\frac{\Delta y}{\Delta \log(x_1)} = \beta_1$$

$\Rightarrow$

$$\frac{\Delta y}{100\Delta \log(x_1)} = \frac{\beta_1}{100}. \tag{11}$$

- Using the approximation

$$100\Delta \log(x) \approx \%\Delta x,$$

we may rewrite (11) as

$$\frac{\Delta y}{\%\Delta x_1} \approx \frac{\beta_1}{100}. \tag{12}$$

- Equation (12) states that in the level-log model

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + u \tag{10}$$

  the quantity

$$\frac{\beta_1}{100}$$

  measures the approximate change in the **level** of y in response to a **one percent** change in the **level** of x_1, holding x_2 constant.

- Therefore, in the level-log model

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + u \tag{10}$$

the predicted change in the value of y in response to a one percent change in x_1, holding x_2 fixed, is given by

$$\frac{\widehat{\beta}_1}{100}.$$

- Consider the level-log regression equation given by

$$bwght = \beta_0 + \beta_1 finc + \beta_2 educ + \beta_3 cigs + u, \tag{13}$$

where *bwght* denotes birth weight of a new born baby in kg, finc denotes family income in dollars, educ denotes the mother's education in years and cigs denotes the number of cigarettes smoked per week by the mother during pregnancy.

- Based on the specification of the regression equation in (13), we predict that, controlling for education and cigarette consumption, a **one dollar** increase in family income increases the average birth weight of a new born baby by β_1 kg.
- However, estimating the impact on birth weight of a one dollar increase in family income is not very meaningful.
- It would be more meaningful to measure the impact on birth weight of a 1% increase in family income.
- Therefore, a more useful specification of the birth weight regression equation is given by

$$bwght = \beta_0 + \beta_1 \log(finc) + \beta_2 educ + \beta_3 cigs + u. \tag{14}$$

In (14) the quantity

$$\frac{\beta_1}{100}$$

measures the average change in the average birth weight of a new born baby associated from a **one percent** change in family income, controlling for education and mother's cigarette consumption.

- When we estimate (14) by OLS the estimated equation is

$$\widehat{bwght} = 3.220 + 0.050\log(finc) + 0.001educ - 0.013cigs. \quad (15)$$

- Based on the result reported in (15), we predict that, controlling for a mother's education and cigarette consumption, **a one percent** increase in family income increases the average birth weight of a new born baby by

$$\frac{0.050}{100} = 0.0005kg = 0.5grams.$$

- Consider the regression equation

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + u. \tag{16}$$

  This model is called a **log-log model** or, more commonly, a **log-linear model** (because it is linear in $\log(y)$ and $\log(x_1)$).

- Because (16) is linear in the regression coefficients, it can be estimated by OLS.

- Suppose that we have a small change in x_1, holding x_2 constant. Then, it follows from (16) that

$$\Delta \log(y) = \beta_1 \Delta \log(x_1),$$

$\Rightarrow$

$$\begin{aligned} \frac{\Delta \log(y)}{\Delta \log(x_1)} &= \beta_1, \\ \frac{100\Delta \log(y)}{100\Delta \log(x_1)} &= \beta_1, \\ \frac{\%\Delta y}{\%\Delta x_1} &= \beta_1. \end{aligned}$$

Therefore, in the log-linear model given by

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + u, \tag{16}$$

the regression coefficient β_1 measures the **percentage change in the level of** y arising from a **one percent change in the level of** x_1, holding x_2 constant.

- Economists call the percentage change in y arising from a one percent change in x_1, holding x_2 constant, the **elasticity** of y with respect to x_1.
- Consider the log-linear regression equation given by

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 \log(mktval) + \beta_3 tenure + u, \tag{17}$$

where *salary* denotes CEO salary in thousands of dollars, *sales* denotes the firms sales in thousands of dollars, *mktval* denotes the market value of the firm's shares and *tenure* denotes the number of years that the CEO has been in his/her current position.

- When we estimate (17) by OLS, the estimated equation is given by

$$\widehat{\log}(salary) = 4.50 + 0.16 log(sales) + 0.11 log(mktval) + 0.01 tenure. \tag{18}$$

- Based on the results reported in (18) we conclude that:
  - Controlling for market value and tenure, a 1% increase in sales is predicted to increase the average salary of CEOs by 0.16%.
  - Controlling for sales and tenure, a 1% increase in market value is predicted to increase average salary of CEOs by 0.11%.
  - Controlling for sales and market value, an extra year of tenure is predicted to increase the average salary of CEOs by 1%.

- There are several factors one must take into account when deciding whether to include a variable in log or level form in a linear regression model.

1. A variable must have a strictly positive range to be a candidate for logarithmic transformation.
2. When deciding whether to include the level or the log of the dependent variable one must think about the nature of the dependent variable. In particular, is it more useful to estimate level changes or % changes in the dependent variable?
3. When deciding whether to include a regressor in level or log form, one must think about the nature of the regressor. Is it more useful to estimate the marginal effect on the dependent variable of a **one unit** change in the regressor or of a **one percent** change in the regressor?

4. Variables which are already measured in percentages are not logged. For example, we don't log interest rates or unemployment rates, since these variable are already measured as percentages.

- Let the variable y be a power function of the variable x given by

$$y = ax^n,$$

where $x > 0$ and a and n are constants.

- The derivative of y with respect to x, which be denote by $\frac{dy}{dx}$, is

$$\frac{dy}{dx} = nax^{n-1}. \tag{19}$$

- For example, let

$$y = 2x^4.$$

Then, using (19) with

$$a = 2, n = 4,$$

we obtain

$$\frac{dy}{dx} = 8x^3.$$

- We interpret $\frac{dy}{dx}$ as the approximate change in y is response to a "small change" in x.
- The smaller the change in x, the more accurate is the approximation.

- Let the variable y be a power function of two variables, x_1 and x_2, given by

$$y = a_1 x_1^n + a_2 x_2^m, \tag{20}$$

where a_1, a_2, n and m are constants.

- The **partial derivative** of y with respect to x_1, which we denote as $\frac{\partial y}{\partial x_1}$, is

$$\frac{\partial y}{\partial x_1} = n a_1 x_1^{n-1},$$

and the partial derivative of y with respect to x_2, which we denote as $\frac{\partial y}{\partial x_2}$, is

$$\frac{\partial y}{\partial x_2} = m a_2 x_2^{m-1}.$$

- We interpret $\frac{\partial y}{\partial x_1}$ as the approximate change in y arising from a small change in x_1, holding x_2 constant, and we interpret $\frac{\partial y}{\partial x_2}$ as the approximate change in y arising from a small change in x_2, holding x_1 constant.
- Let

$$y = a_1 x_1^n x_2^m.$$

Then,

$$\frac{\partial y}{\partial x_1} = n a_1 x_1^{n-1} x_2^m, \tag{21}$$

and

$$\frac{\partial y}{\partial x_2} = m a_1 x_1^n x_2^{m-1}. \tag{22}$$

- Note from (21) and (22) respectively that when we partially differentiate y with respect to x_1 we treat x_2 as a constant and when we partially differentiate y with respect to x_2 we treat x_1 as a constant.

- In the basic linear regression model that we have studied up to this point we assumed that

$$E(y_i|x_i, x_{i2,...,}x_{ik}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}, \tag{23}$$

implying that

$$\begin{aligned} y_i &= E(y_i|x_i, x_{i2,...,}x_{ik}) + u_i \\ &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_{i.} \end{aligned}$$

- An immediate implication of the specification given by (23) is that

$$\frac{\partial E(y_i|x_i, x_{i2,...,}x_{ik})}{\partial x_{ij}} = \beta_j, j = 1, 2, ..., k. \tag{24}$$

- Equation (24) implies that the marginal effect of x_{ij} on the conditional mean of y_i is independent of the level of x_j.
- For example, the wage equation

$$E(wage_i|educ_i, \exp er_i, IQ_i) = \beta_0 + \beta_1 educ_i + \beta_2 \exp er + \beta_3 IQ_i$$

implies that

$$\frac{\partial E(wage_i|educ_i, \exp er_i, IQ_i)}{\partial \exp er_i} = \beta_2.$$

That is, the effect on average wages of an additional year of experience does not depend on the level of one's experience.

- However, in many applications is makes sense to assume that the marginal effect of a regressor on the conditional mean of y_i varies with the value of the regressor.

- For example, in a wage equation we may wish to allow for the possibility that the marginal effect of experience on average wages is different for the population of individuals with 5 years of experience is different from the marginal effect for the population of individuals who have 15 years of experience.
- We can allow for the possibility that the marginal effect of a regressor on the conditional mean of the dependent variable depends on the value of the regressor by making the conditional mean a **quadratic function** of the regressor.

- For example, if we assume that

$$E(y_i|x_{i1}, x_{i2}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i2}, \tag{25}$$

then

$$\begin{aligned}
\frac{\partial E(y_i|x_{i1}, x_{i2})}{\partial x_{i1}} &= \frac{\partial(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i2})}{\partial x_{i1}} \\
&= \frac{\partial(\beta_1 x_{i1} + \beta_2 x_{i1}^2)}{\partial x_{i1} if} \\
&= \frac{\partial \beta_1 x_{i1}}{\partial x_{i1}} + \frac{\partial \beta_2 x_{i1}^2}{\partial x_{i1}} \\
&= \beta_1 + 2\beta_2 x_{i1}. \qquad (26)
\end{aligned}$$

- We see from (26) that the marginal effect of x_{i1} on $E(y_i|x_{i1}, x_{i2})$ depends on the value of x_{i1}.

- For example, when

$$
\begin{aligned}
x_{i1} &= 1, \frac{\partial E(y_i|x_{i1}, x_{i2})}{\partial x_{i1}} = \beta_1 + 2\beta_2 x_{i1} = \beta_1 + 2\beta_2, \\
x_{i1} &= 2, \frac{\partial E(y_i|x_{i1}, x_{i2})}{\partial x_{i1}} = \beta_1 + 2\beta_2 x_{i1} = \beta_1 + 4\beta_2.
\end{aligned}
$$

- When we write

$$E(y_i|x_{i1}, x_{i2}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i2} \tag{25}$$

the associated linear regression equation is

$$
\begin{aligned}
y_i &= E(y_i|x_{i1}, x_{i2}) + u_i \\
&= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i2} + u_i. \qquad (27)
\end{aligned}
$$

- Equation (27) implies that

$$\frac{\partial y_i}{\partial x_{i1}} = \beta_1 + 2\beta_2 x_{i1}.$$

- Even though the regression equation given in (27) contains a quadratic term in x_{i1}, it can still be estimated by OLS because the equation is linear in the regression coefficients.
- **Example:** A researcher who wished to estimate how many minutes on average women sleep in a week specified the following regression equation:

$$sleep_i = \beta_0 + \beta_1 age_i + \beta_2 age_i^2 + \beta_3 educ_i + u_i, 1 = 1, 2, ..., 305. \quad (28)$$

- When (28) was estimated by OLS, the following results were obtained:

$$\widehat{sleep_i} = 4428.07 - 49.30 age_i + 0.58 age_i^2 - 13.92 educ_i. \tag{29}$$

- Based on the results reported in (29):
    - $$\widehat{E}(sleep_i | age_i, educ_i) = 4428.07 - 49.30 age_i + 0.58 age_i^2 - 13.92 educ_i. \tag{30}$$
    - $$\begin{aligned} \frac{\partial \widehat{sleep_i}}{\partial age_i} &= \widehat{\beta}_1 age_i + 2\widehat{\beta}_2 age_i \\ &= -49.30 + 2(0.58) age_i \\ &= -49.30 + 1.16 age_i. \end{aligned} \tag{31}$$

- It follows from (31) that

$$\frac{\partial \widehat{sleep_i}}{\partial age_i} = \begin{cases} 0 \text{ when } age_i = \frac{49.30}{1.16} = 42.5 \\ > 0 \text{ when } age_i > \frac{49.30}{1.16} = 42.5 \\ < 0 \text{ when } age_i < \frac{49.30}{1.16} = 42.5. \end{cases},$$

- Therefore, the predicted average number of minutes slept per week by women decreases up to age 42.5, reaches a minimum at age 42.5 and increases after age 42.5.
- When deciding whether or not to include quadratic terms in a regression equation, the following considerations are relevant:
  1. Does it make more sense to assume that the marginal effect of x_j on the dependent variable depends on the level of x_j or does it make mor sense to assume that it is independent of the level of x_j?
  2. Is x_j^2 statistically significant in the estimated regression equation?

- When specifying a linear regression model we pursue two potentially conflicting goals.
  - We attempt to construct regression models which do a good job of explaining variations in the value of the dependent variable.
  - We attempt to construct regression models which do not contain a very large number of regression coefficients, since each of these regression coefficients has to be estimated using the available sample.
- Ideally, we would like to be able to "explain a lot with a little".
- An econometric model which does not contain "too many" parameters is called a **parsimonious** model.
- The goals of parsimony and predictive power are in conflict because we can almost always increase the predictive power of an econometric model by increasing the number of explanatory variables.

- In this section we briefly consider a variety of criteria which have been proposed for choosing between competing regression models.
- These criteria all involve making trade-offs between specifying a model which is parsimonious and specifying one which does a good job of predicting the values of the dependent variable.

- Recall that

$$R^2 = \frac{SSE}{SST}$$

  is a statistic which measures the proportion of the variation in the dependent variable that is explained or predicted by the explanatory variables.
- While R^2 is a useful descriptive statistic for a single regression, it is of limited use for comparing alternative linear regression models.
- This is because R^2 never decreases and almost always increases each time a new explanatory variable is added to a regression model, no matter how little predictive power the new regressor may have!

- To see this note that

$$
\begin{aligned}
R^2 &= \frac{SSE}{SST} \\
&= \frac{SST - SSR}{SST} \\
&= 1 - \frac{SSR}{SST} \\
&= 1 - \frac{SSR/(n-1)}{SST/(n-1)}.
\end{aligned}
$$

When we add an additional explanatory variable to a linear regression equation, SSR never increases and almost always decreases.

- Therefore,

$$R^2 = 1 - \frac{SSR/(n-1)}{SST/(n-1)} \tag{32}$$

  never decreases and almost always increases.
- Consequently, a linear regression equation with a larger number of explanatory variables will almost always have a larger R^2 than one with fewer explanatory variables.
- For this reason, R^2 cannot be used to compare linear regression models which have a different number of explanatory variables.
- An additional problem with using R^2 to choose between different linear regression models is that it cannot be used to choose between linear regression models which have different dependent variables.

- For example, we cannot use R^2 to choose between the linear regression model

$$Log(wage_i) = \beta_0 + \beta_1 educ_i + u_i$$

and the linear regression model

$$wage_i = \beta_0 + \beta_1 educ_i + u_i,$$

even though they have the same number of explanatory variables.

- An alternative statistic to R^2, called **adjusted** R^2, has been proposed for choosing between linear regression models which have a different number of explanatory variables.
- Unlike R^2, $\overline{R}^2$ does not necessarily increase as the number of explanatory variables increases.

- Adjusted R^2, which we denote by $\overline{R}^2$ (sometimes called R-bar squared), is defined as

$$\overline{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)}. \tag{33}$$

- Comparing (33) and

$$R^2 = 1 - \frac{SSR/(n-1)}{SST/(n-1)}, \tag{32}$$

  we see that the difference between the two statistics is the inclusion of k in the divisor of SSR in the definition of $\overline{R}^2$.

- It is evident from (33) that $\overline{R}^2$ does not necessarily increase when we add an additional explanatory variable.

- This is because although SSR (almost always) decreases when we add an additional explanatory variable to the regression equation, k also increases and the net effect on $\overline{R}^2$ may be either positive or negative.
- It is clear from

$$\overline{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)} \tag{33}$$

  that $\overline{R}^2$ will only increase if the reduction in SSR is sufficiently large to offset the increase in k.
- Given that SSR measures the variation in the dependent variable that is not explained by the regressors, a significant reduction in SSR will only occur if the additional regressor substantially increases the ability of the model to explain the variation in the dependent variable.
- If the additional explanatory variable has very little predictive power, so that its inclusion reduces SSR by a small amount, then $\overline{R}^2$ will fall.

- When we use $\overline{R}^2$ to choose between regression models with the same dependent variable, we choose the model with the highest $\overline{R}^2$.
- While $\overline{R}^2$ can be used to compare linear regression models which have a different number of explanatory variables, it cannot be used to compare linear regression models which have a different dependent variable.
- Note from the definition

$$\overline{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)}. \tag{33}$$

it is possible for

$$\overline{R}^2 < 0$$

if SSR is very large.

- When we estimate a linear regression equation in Eviews, $\overline{R}^2$ is automatically reported.
- For example, if Figure 1 below

$$\overline{R}^2 = 0.159545.$$

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

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5.198085	0.121543	42.76759	0.0000
EDUC	0.057108	0.007348	7.771960	0.0000
EXPER	0.019525	0.003244	6.018132	0.0000
IQ	0.005786	0.000980	5.905770	0.0000

R-squared	0.162244	Mean dependent var	6.779004
Adjusted R-squared	0.159545	S.D. dependent var	0.421144
S.E. of regression	0.386089	Akaike info criterion	0.938773
Sum squared resid	138.7795	Schwarz criterion	0.959481
Log likelihood	-434.8764	Hannan-Quinn criter.	0.946669
F-statistic	60.10079	Durbin-Watson stat	1.811646
Prob(F-statistic)	0.000000		

Figure: 1

- There are three other commonly used criteria for choosing between alternative linear regression models. These three criteria are collectively called **information criteria**.
- Each criterion seeks to achieve a compromise between specifying a model which fits the data well on the one hand and specifying a model which is parsimonious on the other.
- Each information criterion has the form

$$IC = c + \ln(SSR) + \frac{P(k)}{n}, \tag{34}$$

where c is a constant, n is the sample size, SSR is the sum of squared residuals from the estimated regression equation and P(k) is a **penalty term** which is increasing in k, the number of explanatory variables.

- Adding additional explanatory variables (almost always) reduces SSR, which decreases the value of IC, but also increases P(k), which increases the value of IC.
- When using a particular IC to choose between regression models, **the preferred regression model is the one with the smallest value of the IC.**
- The three most commonly used IC are:

1. The Akaike Information Criterion (AIC)

$$AIC = c_1 + \ln(SSR) + \frac{2k}{n}. \tag{35}$$

2. The Hannan-Quinn Criterion (HQ)

$$HQ = c_2 + \ln(SSR) + \frac{2k\ln[\ln(n)]}{n}. \tag{36}$$

3. The Schwarz Information Criterion, which is also known as the Bayes Information Criterion, (SIC/BIC)

$$SIC\,/\,BIC = c_3 + \ln(SSR) + \frac{k\ln(n)}{n}. \tag{37}$$

- Comparing (35), (36) and (37) to the general representation for an IC given by

$$IC = c + \ln(SSR) + \frac{P(k)}{n}, \tag{34}$$

it is evident that for the

$$\begin{aligned} AIC &: \quad P = 2k \\ HQ &: \quad P = 2k\ln[\ln(n)] \\ SIC\,/\,BIC &: \quad P = k\ln(n). \end{aligned}$$

- For
$$n > 16$$
we have the following ranking of the penalties imposed by the three IC:
$$P(SIC/BIC) > P(HQ) > P(AIC). \quad (38)$$
- It follows from (38) that the SIC/BIC penalizes additional regressors more severely than the HQ, which in turn penalizes additional regressors more severely than the AIC.
- Consequently, the regression model chosen by the SIC/BIC will be at least as parsimonious as the model chosen by the HQ, which in turn will be at least as parsimonious as that chosen by the AIC.

- Recall that when we estimated

$$Log\left(wage_i\right) = \beta_0 + \beta_1 educ_i + \beta_2 \exp er + \beta_3 IQ_i + u_i, i = 1, 2, ...935 \tag{9}$$

by OLS we obtained the output reported in Figure 1 below.

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

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5.198085	0.121543	42.76759	0.0000
EDUC	0.057108	0.007348	7.771960	0.0000
EXPER	0.019525	0.003244	6.018132	0.0000
IQ	0.005786	0.000980	5.905770	0.0000

R-squared	0.162244	Mean dependent var	6.779004
Adjusted R-squared	0.159545	S.D. dependent var	0.421144
S.E. of regression	0.386089	Akaike info criterion	0.938773
Sum squared resid	138.7795	Schwarz criterion	0.959481
Log likelihood	-434.8764	Hannan-Quinn criter.	0.946669
F-statistic	60.10079	Durbin-Watson stat	1.811646
Prob(F-statistic)	0.000000		

Figure: 1

- Notice that the AIC,HQ and SIC/BIC for this regression equation are automatically reported by Eviews.

$$\begin{aligned} AIC &= 0.938773, \\ SIC\,/\,BIC &= 0.959481, \\ HQ &= 0.946669. \end{aligned}$$

- Information criteria such as the AIC,HQ and SIC/BIC are most frequently used to choose between alternative specifications for pure time series models such as

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + ... + \beta_p y_{t-p} + u_t. \tag{39}$$

Special cases of (39) are

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + u_t, \tag{40}$$

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} + u_t. \tag{41}$$

We could use an information criterion to choose between (40) and (41).

- Unfortunately, different information criteria often choose different models and when this occurs it is unclear which information criterion to rely on.