In the previous chapter we specified the basic linear regression model and distinguished between the population regression and the sample regression.

Our objective is to make use of the sample data on Y and X and obtain the "**best**" estimates of the population parameters.

The most commonly used procedure used for regression analysis is called **ordinary least squares** (**OLS**).

The OLS procedure minimizes the sum of squared residuals.

From the theoretical regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

we want to obtain an estimated regression equation

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

OLS is a technique that is used to obtain $\hat{\beta}_0$ and $\hat{\beta}_1$. The OLS procedure minimizes

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$. Solving the minimization problem results in the following expressions:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Notice that different datasets will produce different values for $\hat{\beta}_0$ and $\hat{\beta}_1$.

Why do we bother with OLS?

Example

Let’s consider the simple linear regression model in which the price of a house is related to the number of square feet of living area (SQFT).

Dependent Variable: PRICE
Method: Least Squares
Sample: 1 14
Included observations: 14

Variable	Coefficient	Std. Error	t-Statistic	Prob.
SQFT	0.138750	0.018733	7.406788	0.0000
C	52.35091	37.28549	1.404056	0.1857

R-squared	0.820522	Mean dependent var	317.4929
Adjusted R-squared	0.805565	S.D. dependent var	88.49816
S.E. of regression	39.02304	Akaike info criterion	10.29774
Sum squared resid	18273.57	Schwarz criterion	10.38904
Log likelihood	-70.08421	F-statistic	54.86051
Durbin-Watson stat	1.975057	Prob(F-statistic)	0.000008

For the general model with k independent variables:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i,$$

the OLS procedure is the same. We choose the $\hat{\beta}$s that minimize the sum of squared residuals.

We let EViews do this for us.

Example

Suppose we would like to include more home characteristics in our previous example. Besides the square footage, price is related to the number of bathrooms as well as the number of bedrooms.

Dependent Variable: PRICE
Method: Least Squares
Sample: 1 14
Included observations: 14

Variable	Coefficient	Std. Error	t-Statistic	Prob.
SQFT	0.154800	0.031940	4.846516	0.0007
BEDRMS	-21.58752	27.02933	-0.798670	0.4430
BATHS	-12.19276	43.25000	-0.281913	0.7838
C	129.0616	88.30326	1.461573	0.1746

R-squared	0.835976	Mean dependent var	317.4929
Adjusted R-squared	0.786769	S.D. dependent var	88.49816
S.E. of regression	40.86572	Akaike info criterion	10.49342
Sum squared resid	16700.07	Schwarz criterion	10.67600
Log likelihood	-69.45391	F-statistic	16.98894
Durbin-Watson stat	1.970415	Prob(F-statistic)	0.000299

Overall Goodness of Fit

No straight line we estimate is ever going to fit the data perfectly. We thus need to be able to judge how much of the variation in Y we are able to explain by the estimated regression equation.

The amount of variation to be explained by the regression is

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

This is referred to as the **total sum of squares** (**TSS**).

Now, we can re-write $Y_i - \bar{Y}$ as

$$Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

From this we can re-write the total variation as follows

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} e_i^2$$

The **explained sum of squares** (**ESS**) represents the explained variation. ESS measures the total variation of $\hat{Y}_i$ from $\bar{Y}$.

The **residual sum of squares** (**RSS**) represents the unexplained variation.

The ratio

$$\frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} = \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

is called the **coefficient of determination** and is denoted by R^2. R^2 lies between 0 and 1.

R^2 measures the percentage of variation of Y around $\bar{Y}$ that is explained by the regression equation. The closer the observed points are to the estimated regression line, the better the fit, the higher the R^2.

The way we have defined R^2 is problematic. The addition of any X variable, will never decrease the R^2. In fact, R^2 is likely to increase.

A different measure of goodness of fit is used, the **adjusted $\boldsymbol{R^2}$** (or **R-bar squared**):

$$\bar{R}^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2/(n-k-1)}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2/(n-1)}$$

This value has a maximum of 1 but a minimum that can be negative.

In addition to the R^2, there is the **simple correlation coefficient**, r, which measures strength and direction of a linear relationship between two variables:

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}},$$

r lies between –1 and 1.