CHULALONGKORN UNIVERSITY
FACULTY OF ECONOMICS

2946653 Research Methods in Labour Economics
and Human Resource Management

EXERCISE 2

Two-Variable Regression Model

Let's journey back to the lessons you have learnt in the Quantitative Methods class on regression analysis. Let Y be the dependent variable while X be the independent or explanatory variable. In Layman's term, we explain Y through X. We can write the **population regression function** (PRF) as the conditional expectation of Y on X as

$$E\left(Y|X_i\right) = f\left(X_i\right)$$

where $f\left(X_i\right)$ denotes some function of X. Assume that $f\left(X_i\right)$ takes a linear form of

$$E\left(Y|X_i\right) = \beta_0 + \beta_1 X_i$$

where β_0 and β_1 are the intercept and slope coefficients of this regression function. As you may notice, this linear functional form may not fitted very well. There may be some deviation between the actual Y_i and the conditional expectation $E\left(Y|X_i\right)$. Write this deviation or stochastic error as

$$\begin{aligned} \varepsilon_i &= Y_i - E\left(Y|X_i\right) \text{ or} \\ Y_i &= \beta_0 + \beta_1 X_i + \varepsilon_i \end{aligned}$$

In reality, knowing the population of Y and X are impossible. In fact, we have a sample of Y corresponding to some fixed value of X to work on. Analogous to the population regression function, we can write the **sample regression function** (SRF) as follow

$$Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\varepsilon}_i$$

where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are estimators of β_0 and β_1 respectively. Our important task is to find the best way to estimate β_0 and β_1. It can be done through the *least-squares* method, i.e., we minimise

$$\begin{aligned} \sum_i \widehat{\varepsilon}_i &= \sum_i \left(Y_i - \widehat{Y}_i\right)^2 \\ &= \sum_i \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i\right)^2 \end{aligned}$$

The estimators we have from this methods are known as the least-squares estimators and their expressions are

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(X_i - \overline{X}\right)^2} \\ \widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{X} \end{aligned}$$

We refer to $\widehat{\beta}_0$ and $\widehat{\beta}_1$ as the OLS estimators of β_0 and β_1.There are several important assumptions regarding the properties and behaviours of X and ε. These assumptions are also known as the *classical assumptions*.

Assumption 1 X is nonstochastic, i.e., X values are fixed.

Assumption 2 The disturbance or error term has zero mean.

$$E\left(\varepsilon_i|X_i\right) = 0 \text{ for all } i$$

Assumption 3 Homoscedasticity of the error terms.

$$E\left(\varepsilon_i^2|X_i\right) = \sigma^2 \text{ for all } i$$

Assumption 4 No autocorrelation between the error terms.

$$E\left(\varepsilon_i, \varepsilon_j\right) = 0 \text{ for } i \neq j$$

Assumption 5 Zero covariance between the error term and the explanatory variable.

$$E\left(\varepsilon_i, X_i\right) = 0 \text{ for all } i$$

Assumption 6 The regression model is correctly specified.

Without walking you through tedious derivation, the variances and standard errors of the OLS estimators are

$$\begin{aligned} var(\widehat{\beta}_1) &= \frac{\sigma^2}{\sum\left(X_i - \overline{X}\right)^2} \\ se(\widehat{\beta}_1) &= \frac{\sigma}{\sqrt{\left(X_i - \overline{X}\right)^2}} \\ var(\widehat{\beta}_0) &= \frac{\sum X_i^2}{N\sum\left(X_i - \overline{X}\right)^2}\sigma^2 \\ se(\widehat{\beta}_0) &= \sigma\sqrt{\frac{\sum X_i^2}{N\sum\left(X_i - \overline{X}\right)^2}} \end{aligned}$$

where var =variance, se =standard error (or standard deviation), σ^2 = the 'true' variance of ε_i and N =number of observations. Since σ^2 is unknown, we can calculate the OLS estimator of σ^2 by the following formula

$$\widehat{\sigma}^2 = \frac{\sum \widehat{\varepsilon}_i^2}{n-2}$$

where

$$\sum \widehat{\varepsilon}_i^2 = \sum_i \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i\right)^2$$

as stated before.

Given that all the classical assumptions are satisfied, the OLS estimators are said to be **BLUE** (Best Linear Unbiased Estimators), i.e., they are unbiased and have minimum variance. This is the well-known **Gauss-Markov Theorem**.

Our next task is to determine whether our sample regression line fits the data, i.e., we want to determine the **goodness of fit**. This can be measured through the **coefficient of determination** or R^2. First, we start from the derivation of the deviation form of the sample regression function. From

$$Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\varepsilon}_i$$

take summation on both sides

$$\sum Y_i = n\widehat{\beta}_0 + n\widehat{\beta}_1 \sum X_i + \sum \widehat{\varepsilon}_i$$

Note that $\sum \widehat{\varepsilon}_i = 0$. Divide both sides by N,

$$\overline{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{X}_i$$

Subtracting the above equation from $Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\varepsilon}_i$,

$$Y_i - \overline{Y}_i = \widehat{\beta}_1 (X_i - \overline{X}_i) + \widehat{\varepsilon}_i$$

Define $y_i = Y_i - \overline{Y}_i$ and $x_i = X_i - \overline{X}_i$, we can write the sample regression function in the *deviation form* as follow:

$$y_i = \widehat{\beta}_1 x_i + \widehat{\varepsilon}_i$$

Recall that

$$Y_i = \widehat{Y}_i + \widehat{\varepsilon}_i$$

Or in the deviation form

$$y_i = \widehat{y}_i + \widehat{\varepsilon}_i$$

(Try to derive this deviation form. It is not difficult.) Squaring on both sides and taking summation,

$$\sum y_i^2 = \widehat{\beta}_1^2 \sum x_i^2 + \sum \widehat{\varepsilon}_i$$

We define $\sum y_i^2$ = the total sum of squares (TSS), $\widehat{\beta}_1^2 \sum x_i^2$ = the explained sum of squares (ESS), and $\sum \widehat{\varepsilon}_i$ = the residual (or unexplained) sum of squares (RSS),

$$TSS = ESS + RSS$$

i.e., total variation = variation due to regression (ESS) + variation due to residual (RSS). Divide both sides by TSS,

$$\frac{ESS}{TSS} + \frac{RSS}{TSS} = 1$$

Define

$$\frac{ESS}{TSS} = R^2 = \frac{\sum \widehat{y}_i^2}{\sum y_i^2}$$

or, alternatively,

$$\begin{aligned} R^2 &= 1 - \frac{RSS}{TSS} \\ &= 1 - \frac{\sum \widehat{\varepsilon}_i}{\sum y_i^2} \end{aligned}$$

Intuitively, R^2 measures the percentage of the total variation in Y explained by the regression model. Three more alternative formulas of R^2 are

$$\begin{aligned} R^2 &= \widehat{\beta}_1^2 \frac{\sum x_i^2}{\sum y_i^2} \text{ or} \\ &= \widehat{\beta}_1^2 \frac{\widehat{var}\,(X)}{\widehat{var}(Y)} \text{ or} \\ &= \frac{(\sum x_i y_i)^2}{\sum x_i^2 \sum y_i^2} \end{aligned}$$

Now you have the basic concepts of the simple regression analysis, it will not be difficult for you to move from the two-variable case to the multiple regression analysis. You can find the derivation of the OLS estimators, their variances, and the coefficient of determination in any econometric textbook.

REMARK

Do NOT forget to save your work. We will come back to use them!

Question 1: Simple Regression Model

Consider data sets in `dataset2.xls`. There are 5 series of data. Consider the following models

$$\begin{aligned}
\text{Model 1} &: V_{1i} = \alpha_0 + \alpha_1 V_{2i} + \varepsilon_i \\
\text{Model 2} &: V_{1i} = \beta_0 + \beta_1 V_{3i} + \varepsilon_i \\
\text{Model 3} &: V_{1i} = \gamma_0 + \gamma_1 V_{4i} + \varepsilon_i \\
\text{Model 4} &: V_{1i} = \delta_0 + \delta_1 V_{5i} + \varepsilon_i
\end{aligned}$$

1. Find the OLS estimators for all models
2. Give the interpretation of your OLS estimators in (1)
3. Find the variances, standard deviations, and covariances of all the OLS estimators
4. Find the coefficients of determination for all models
5. Which model is the best among these four? Why?
6. Now consider

$$\text{Model 5} : V_{1i} = \theta_0 + \theta_1 V_{2i} + \theta_2 V_{3i} + \theta_3 V_{4i} + \theta_4 V_{5i} + \varepsilon_i$$

   Is the model 5 better that the above four? Explain.

7. Consider

$$V_{1i} = \phi_0 + \phi_1 V_{2i} + \phi_2 2V_{2i} + \varepsilon_i$$

   Find the OLS estimators of this model. Comment on your findings.

Question 2: Exchange Rate and Imports

Consider data in `dataset3.xls`. Definitions of the variables are

$$\begin{aligned}
EXUS &= \text{exchange rate (baht/US\$)} \\
BENZ &= \text{quality sales of benzines (million litres)} \\
IMNDG &= \text{import of non-durable consumption goods (million US\$)} \\
IMDG &= \text{import of durable consumption goods (million US\$)}
\end{aligned}$$

(Data Source: Bank of Thailand and NESDB)

1. Find the OLS estimators from the following model and interpret $\widehat{\alpha}_1$

$$\text{Model 1} : EXUS_t = \alpha_0 + \alpha_1 BENZ_t + \varepsilon_t$$

2. Find the OLS estimators from the following model and interpret $\widehat{\beta}_1$

$$\text{Model 2} : BENZ_t = \beta_0 + \beta_1 EXUS + \varepsilon_t$$

3. Comment on Model 1 and 2. Which model will you use? Why?

4. Consider the models

$$\begin{aligned} \text{Model 3}: &\quad BENZ_t = \gamma_0 + \gamma_1 IMNDG_t + \varepsilon_t \\ \text{Model 4}: &\quad BENZ_t = \delta_0 + \delta_1 IMDG_t + \varepsilon_t \end{aligned}$$

Find the OLS estimators. Which variables between $IMNDG$ and $IMDG$ will you choose in order to explain $BENZ$? Why? Also, between Model 3 and 4, which model makes more sense? Give your reasons.

5. From Model 5 and 6

$$\begin{aligned} \text{Model 5} &: \quad IMNDG_t = \eta_0 + \eta_1 EXUS_t + \varepsilon_t \\ \text{Model 6} &: \quad IMDG_t = \theta_0 + \theta_1 EXUS_t + \varepsilon_t \end{aligned}$$

Find and interpret the OLS estimators.

Question 3: Liquor, Electricity, GDP, and Exchange Rate

From `dataset4.xls`, you are asked to construct two regression models. Clearly write down your models. One of them has to be multiple regression model. Explain the reasons behind your choice of dependent and explanatory variables. Guess the sign of the coefficients you are about to estimate. Find and interpret the OLS estimators of your models. Are all the signs consistent with your guess?

Variables in `dataset4.xls` are

$LIQUOR$ = liquor consumption (20,000 litres)

$BEER$ = beer consumption (10,000 litres)

$SODA$ = soda and softdrink consumption (million litres)

GDP = Gross Domestic Product (current price) (million baht)

$ELECHH$ = household consumption of electricity (million-kilowatt/hour)

$EXUS$ = exchange rate (baht/US$)

(Data Source: Bank of Thailand and NESDB)

Apart from these six variables, what other variables do you want to add to your model? Why?

Question 4: The Relationship between Earnings and Education

Let Y be average hourly earnings (US$) and X be years of schooling. We regress Y on X and obtain the following results.

$$\begin{aligned} \widehat{Y_i} &= 0.7437 + 0.6416 X_i \\ R^2 &= 0.8944 \end{aligned}$$

Interprete these results.

Question 5: Consumption-Income Relationship

We estimate the Keynesian cunsumption function using the United States data during 1982-1996, all measured in 1992 billions of dollars. We regress personal consumption expenditure (PCE) on Gross Domestic Product (GDP). The findings are

$$\begin{aligned} \widehat{PCE_t} &= -184.0780 + 0.7064 GDP_t \\ R^2 &= 0.9984; var(\widehat{\beta}_0) = 2140.17.7; var(\widehat{\beta}_1) = 0.000061 \end{aligned}$$

Do the sign of both coefficients make any sense? Is there any economic reason behind these results? Is there anything suspicious about these results?