

Chapter 2: Ordinary Least Squares

Lecture 2

Revised on October 12, 2007

1 Multiple Regression

1.1 The Population Regression

When we study demand for a good from households, we often consider income of each household in addition to the price of the good.

If there are additional independent variables, the population regression is

$$E(Y_i|X_{1i}, \dots, X_{Ki}) = \beta_0 + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} \quad (1)$$

For example, for $K = 2$,

$$E(Y_i|X_{1i}, X_{2i}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \quad (2)$$

The coefficients, $\beta_1, \beta_2, \dots, \beta_K$ are called *partial regression coefficients*. Let $\epsilon_i = Y_i - E(Y_i|X_{1i}, \dots, X_{Ki})$ be the stochastic error term. Then

$$Y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} + \epsilon_i \quad (3)$$

where the systematic component is $\beta_0 + \beta_1 X_{1i} + \dots + \beta_K X_{Ki}$, which is $E(Y_i|X_{1i}, \dots, X_{Ki})$, and the nonsystematic component is ϵ_i .

1.2 The meaning of partial regression coefficient

In Equation (2), β_1 measures the change in the expected value of Y per unit change in X_1 , holding the value of X_2 constant. Likewise, β_2 measures the change in the expected value of Y per unit change in X_2 , holding the value of X_1 constant.

1.3 Estimating Multiple Regressions Models with OLS

For the regression model in Equation (2), the least squares principle tells us to choose $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$, so that we minimize residual sum of squares (RSS).

$$\begin{aligned}
RSS(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) &= \sum_{i=1}^N e_i^2 = \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 \\
&= \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \beta_2 X_{2i})^2. \quad (4)
\end{aligned}$$

For $K > 2$, we also minimize RSS to obtain OLS estimates. The OLS estimators are denoted by $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_K$. The fitted value of sample regression is written as

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \dots + \beta_K \hat{X}_{Ki} \quad (5)$$

By adding the residual, we write

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \dots + \beta_K \hat{X}_{Ki} + e_i \quad (6)$$

See Section 2.2.3 of Studenmund for an example of a multiple regression.

2 Describing the Overall fit of the Estimated Model

As a measure of "goodness of fit" we have the *coefficient of determination*, denoted by the symbol R^2 (read as r squared).

2.1 Total, Explained, and Residual Sums of Squares

In order to define R^2 , we need to develop some measures of variation.

The *total sum of squares*, or TSS, is

$$TSS = \sum_{i=1}^N (Y_i - \bar{Y})^2. \quad (7)$$

This is a measure of variation of the dependent variable, and becomes larger when an observed value of the dependent variable deviate more from its sample average.

The *explained sum of squares*, or ESS, is

$$ESS = \sum_{i=1}^N (\hat{Y}_i - \bar{Y})^2. \quad (8)$$

This is a measure of variation of the fitted value, and becomes larger when the fitted value for i deviates more from the sample average of the dependent variable.

The residual sum of squares, or RSS, is

$$RSS = \sum_{i=1}^N e_i^2. \quad (9)$$

This is a measure of variation of the residual, and becomes larger when the absolute value of the residual for i is larger.

In order to see the relationship between TSS, ESS, and RSS, recall

$$Y_i = \hat{Y}_i + e_i \quad (10)$$

If we subtract \bar{Y} from the both sides of this equation, we obtain

$$(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + e_i \quad (11)$$

Squaring each term and summing over the sample, we obtain

$$\sum_{i=1}^N (Y_i - \bar{Y})^2 = \sum_{i=1}^N (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^N e_i^2. \quad (12)$$

because the cross term of the right hand side is zero when OLS is used to obtain \hat{Y}_i .

Equation (12) decomposes the TSS into two components. One is ESS, and the other is RSS.

$$TSS = ESS + RSS. \quad (13)$$

See Figure 2.3 of Studenmund.

2.2 R^2 , The Coefficient of Determination

We define R^2 by

$$R^2 = \frac{ESS}{TSS} \quad (14)$$

Note that $0 \leq R^2 \leq 1$, and that a convenient formula to compute R^2 is

$$R^2 = 1 - \frac{RSS}{TSS} \quad (15)$$

A value of R^2 close to one shows an excellent overall fit, whereas a value near zero shows a failure of the estimated regression equation to explain the values of Y_i better than could be explained by the sample average \bar{Y} . R^2 measures the fraction of the variation of Y around \bar{Y} that is explained by the estimated regression equation.

2.3 Adjusted R^2

Because RSS decreases as you add more independent variables, that R^2 always increases when we add an additional explanatory variable. This does not necessarily mean that adding a variable is good.

Example Let Y be the weight and X_1 be height over five feet in Studentmund's weight guessing regression in Section 1.4. Then

$$\begin{aligned}\hat{Y}_i &= 103.40 + 6.38X_{1i} \\ N &= 20, R^2 = 0.74\end{aligned}\tag{16}$$

Imagine that you add a completely nonsensical variable (say, the campus post office box number of each individual) as X_2 , and that the results become

$$\begin{aligned}\hat{Y}_i &= 103.35 + 6.38X_{1i} + 0.02X_{2i} \\ N &= 20, R^2 = 0.75\end{aligned}\tag{17}$$

Here, no matter how nonsensical the additional variable is, R^2 cannot decrease as long as you use OLS.

The adjusted R^2 , denoted by the symbol \bar{R}^2 is defined by

$$\bar{R}^2 = 1 - (1 - R^2) \frac{N - 1}{N - K - 1} \quad (18)$$

Exercise: Compute the adjusted R^2 for the two regressions in the example for the weight guessing game above. Which regression do you prefer? Why?