7. Multiple Regression Analysis: The Problem of Analysis

Three-Variable Model: Notation and Assumptions

Let us consider the following three-variable PRF as:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

where

Y_i is the dependent variable (regressand)

X_{2i} and X_{3i} are the regressors or the explanatory variables

u_i is the stochastic disturbance term

Remark: the subscript i is denoted the observation i from our sample data.

In case our data are time series, the subscript t will denote the t observation.

β_1 means the average value of Y when X_2 and X_3 are set equal to zero

β_2 and β_3 are called the partial regression coefficients.

We will talk about the meaning of β_1 and β_2 shortly after knowing the assumptions of the classical linear regression model (CLRM)

Page 2

Tuesday, April 10, 2018 7:52 AM

Under the CLRM, we assume:

1. Zero mean value of u_i

2. No serial correlation

3. Homoscedasticity

4. Zero covariance between u_i and each X variable, or

5. No specification bias or

The model is correctly specified.

6. No exact collinearity between the X variables or

Page 3

Tuesday, April 10, 2018 7:52 AM

By the above assumptions, we can find out the conditional expectation of Y_i:

The meaning of partial coefficients:

β_2

β_3

7.1 OLS Estimation of the Partial Regression Coefficients

In order to find the OLS estimators, we need to write down the sample regression function (SRF) corresponding to the PRF:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{u}_i$$

Page 4

Tuesday, April 10, 2018 7:52 AM

From the FOC, we then get the normal equations:

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}_2 + \hat{\beta}_3 \bar{X}_3$$
$$\sum Y_i X_{2i} = \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 + \hat{\beta}_3 \sum X_{2i} X_{3i}$$
$$\sum Y_i X_{3i} = \hat{\beta}_1 \sum X_{3i} + \hat{\beta}_2 \sum X_{2i} X_{3i} + \hat{\beta}_3 \sum X_{3i}^2$$

We therefore get:

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3$$
$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$$
$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$$

Variance and Standard Errors of OLS Estimators

$$var(\hat{\beta}_1) = \left[\frac{1}{n} + \frac{\bar{X}_2^2 \sum x_{3i}^2 + \bar{X}_3{}^2 \sum x_{2i}^2 - 2\bar{X}_2 \bar{X}_3 \sum x_{2i} x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - (\sum x_{2i} x_{3i})^2}\right] * \sigma^2$$
$$se(\hat{\beta}_1) = +\sqrt{var(\hat{\beta}_1)}$$

$$var(\hat{\beta}_2) = \frac{\sum x_{3i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} * \sigma^2$$
$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)}$$
$$se(\hat{\beta}_2) = +\sqrt{var(\hat{\beta}_2)}$$

$$var(\hat{\beta}_3) = \frac{\sum x_{2i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} * \sigma^2$$
$$var(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2 (1 - r_{23}^2)}$$
$$se(\hat{\beta}_3) = +\sqrt{var(\hat{\beta}_2)}$$

Page 5

Tuesday, April 10, 2018 7:52 AM

$$cov(\hat{\beta}_2, \hat{\beta}_3) = \frac{-r_{23}\sigma^2}{(1-r_{23}^2)\sqrt{\sum x_{2i}^2}\sqrt{\sum x_{3i}^2}}$$

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-3}$$

7.2 Properties of OLS Estimators

Page 6

Tuesday, April 10, 2018 7:52 AM

Properties of OLS Estimators (Cont:)

Page 7

Tuesday, April 10, 2018 7:52 AM

Properties of OLS Estimators (Cont:)

Page 8

Tuesday, April 10, 2018 7:52 AM

The Multiple Coefficient of Determination R^2 and the Multiple Coefficient of Correlation R

In this section, we will study how to measure the proportion of the variation in Y explained by the variables X_2 and X_3 jointly. This is the same concept of r^2 that we have learned before.

The quantity that gives this information is known as the **the multiple coefficient of determination** and is denoted by R^2.

To derive R^2, we firstly write down the following equation:

$$\begin{aligned} Y_i &= \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{u}_i \\ &= \hat{Y}_i + \hat{u}_i \end{aligned} \tag{7.1}$$

where $\hat{Y}_i$ is the estimated value of Y_i from the fitted regression line and is an estimator of true $E(Y_i|X_{2i}, X_{3i})$.

7.1 may be written as

$$\begin{aligned} y_i &= \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i \\ &= \hat{y}_i + \hat{u}_i \end{aligned} \tag{7.2}$$

Squaring 7.2 on both sides and summing over the sample values, we obtain

$$\boxed{\begin{aligned} \sum y_i^2 &= \sum \hat{y}_i{}^2 + \hat{u}_i{}^2 + 2\sum \hat{y}_i \hat{u}_i \\ &= \sum \hat{y}_i{}^2 + \sum \hat{u}_i{}^2 \end{aligned} \quad (7.3)}$$

Page 9

Tuesday, April 10, 2018 7:52 AM

$$R^2 = \frac{ESS}{TSS} = \frac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}}{\sum y_i^2} \tag{7.4}$$

$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)}$$

$$var(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2 (1 - r_{23}^2)}$$

correlation coefficient between X_{2i} and X_{3i}

so, given $\sum x_{2i}$, $\sum x_{3i}$, and σ^2, $\uparrow r_{23} \rightarrow \uparrow r_{23}^2 \rightarrow \uparrow var(\beta_2)$:(

$\uparrow var(\beta_3)$:(

The three-or-more-variable analogue of r is the coefficient of multiple correlation, denoted by R, and it is a measure of the degree of association between Y and all the explanatory variables jointly. Although r can be positive or negative, R is always taken to be positive.

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2}\left(\frac{1}{1 - R_j^2}\right)$$

R_j^2 is the R^2 in the regression of X_j on the remaining $(k-2)$ regressors

Ex: $X_{2i} = \hat{\alpha}_1 + \hat{\alpha}_2 X_{3i}$

$\Downarrow$

you get $R_{x_2}^2$

7.2.1 R^2 and the Adjusted R^2

It should be noted that he R^2 is a nondecreasing function of the number of explanatory variables. Thus, when the number of regressors increases, R^2 almost invariably increases and never decreases. **In other words, an additional X variable will not decrease R^2!**

To explain this fact, let we write down the definition of R^2 again:

$$
\begin{aligned}
R^2 &= \frac{ESS}{TSS} \\
&= 1 - \frac{RSS}{TSS} \\
&= 1 - \frac{\sum \hat{u}_i{}^2}{\sum y_i^2}
\end{aligned}
\tag{7.5}
$$

$\Sigma \hat{u}_i^2$

Therefore, in comparing two regression models **with the same dependent variable but differing number of X variables**, one should be very wary of choosing the model with the highest R^2.

In light of comparing two R^2 terms, we have to take into account the number of X variables present in the model. To achieve this goal , we can consider the alternative coefficient of determination, which is as follows:

$$\bar{R}^2 = 1 - \frac{\Sigma \hat{u}_i^2 / (n-3)}{\Sigma y_i^2 / (n-1)}$$

for 3-variable regression model ($k=3$)

k= the number of parameters in the model including the intercept term.
n= the number of observations in the sample data.

The above equation is known as **the adjusted** R^2, denoted by $\bar{R}^2$. The term adjusted means adjusted for the df associated with the sums of squares entering into 7.5.

General form for $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_2 X_{3i} + \ldots + \beta_k X_{ki}$

Then $\bar{R}^2 = 1 - \frac{\Sigma \hat{u}_i^2 / (n-k)}{\Sigma y_i^2 / (n-1}$

We can rewrite the the adjusted $\bar{R}^2$ as:

variance of residual term or residual variance

$$\bar{R}^2 = 1 - \frac{\hat{\sigma}^2}{S_Y^2}$$

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-k}$$ (unbiased estimator of σ^2)

$$S_Y^2 = \frac{\sum y_i}{n-1}$$ (sample variance of Y)

We can also get the equation which shows the relationship between $\bar{R}^2$ and R^2:

Recall that $R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2}$ —(1)

$$\bar{R}^2 = 1 - \frac{\sum \hat{u}_i^2/(n-k)}{\sum y_i^2/(n-1)}$$ —(2)

Substituting (1) into (2) gives:

$$\boxed{\bar{R}^2 = 1 - (1 - R^2) \cdot \frac{(n-1)}{n-k}}$$ → verify this @ home.

Besides R^2 and $\bar{R}^2$ as goodness of fit measures, other criteria are often used to judge the adequacy of a regression model. Two of these are **Akaike's Information criterion and Amemiya's Prediction criteria**, which are used to select between competing models. We will discuss these criteria in greater detail later.

1. For $k>1$, $\bar{R}^2 < R^2$.
2. If $k \uparrow$, R^2 will always increase BUT NOT NECESSARILY FOR $\bar{R}^2$!
3. $\bar{R}^2$ can be negative. Ex: when $R^2 = 0$.
4. ** When you compare two $\bar{R}^2$ values, make sure that
   1. same sample size
   & 2. same dependent variable

$$\ln Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$ —(1)

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + \beta_3 X_{3i} + u_i$$ —(2)

$$\beta_2 = \frac{d \ln Y}{d X_i} = \frac{\frac{\Delta Y}{Y}}{\Delta X} \rightarrow 100 \cdot \beta_2 = \frac{\frac{\Delta Y}{Y} \cdot 100}{\Delta X}$$

$$\beta_2 = \frac{d \ln Y}{dx_2} = \frac{\frac{\Delta Y}{Y}}{\Delta X} \rightarrow 100 \cdot \beta_2 = \frac{\frac{\Delta Y}{Y} \cdot 100}{\Delta X}$$

$$\alpha_2 = \frac{dY}{dx_2}$$

Since ① and ② differ in dependent var, $\bar{R}^2$ or R^2 between the two models CANNOT BE DIRECTLY COMPARED!

After songkran, we will learn how to manage this if we want to compare R^2 between model ① and ②.