EC220 classes

November 26, 2008

Antoine Goujard

The dummy variable trap

Let S_i be the number of years of education of an individual, i, and $MALE_i$ and $FEMALE_i$ be two dummy variables taking value 1 if the individual is a male or if the individual is a female. Suppose you want to fit a simple educational attainment function:

(1) $S_i = \beta_1 + \beta_2.MALE_i + u_i$

However you make a mistake and include the two dummy variables for males and females at the same time. In other words, you are trying to fit the following equation:

(2) $S_i = \beta_1 + \beta_2.MALE_i + \beta_3.FEMALE_i + u_i$

Intuition and knowledge of the **dummy variable trap** should warn you that you can not find estimates of β_1, β_2, β_3 in model (2).

Indeed the formula for the estimator of β_2 in model (2) is given in the book p.123 (formula, 3.11):

$b_2 = \frac{\sum_{i=1}^n (X_{2i}-\bar{X}_2)(Y_i-\bar{Y})\sum_{i=1}^n (X_{3i}-\bar{X}_3)^2 - \sum_{i=1}^n (X_{3i}-\bar{X}_3)(Y_i-\bar{Y})\sum_{i=1}^n (X_{2i}-\bar{X}_2)(X_{3i}-\bar{X}_3)}{\sum_{i=1}^n (X_{2i}-\bar{X}_2)^2.\sum_{i=1}^n (X_{3i}-\bar{X}_3)^2 - (\sum_{i=1}^n (X_{2i}-\bar{X}_2)(X_{3i}-\bar{X}_3))^2}$

Or,

$b_2 = \frac{\sum_{i=1}^n (X_{2i}-\bar{X}_2)(Y_i-\bar{Y})\sum_{i=1}^n (X_{3i}-\bar{X}_3)^2 - \sum_{i=1}^n (X_{3i}-\bar{X}_3)(Y_i-\bar{Y})\sum_{i=1}^n (X_{2i}-\bar{X}_2)(X_{3i}-\bar{X}_3)}{\sum_{i=1}^n (X_{2i}-\bar{X}_2)^2.\sum_{i=1}^n (X_{3i}-\bar{X}_3)^2 \times (1-r^2_{X_2X_3})}$

But, $r^2_{FEMALE\,MALE} = 1$

Because, $FEMALE_i = 1 - MALE_i$.

Hence, in this particular case, because of the **exact multi-collinearity between $MALE_i$, $FEMALE_i$ and the constant term**, the estimates can not be computed.

Exact multi-collinearity occurs when one of the explanatory variable (including the constant) is equal to a linear combination of the other explanatory variables. Here the constant is equal to the sum of the dummy "FEMALE" and the dummy "MALE". Note that we have already seen this problem in PS1 item4 (or 1.12 in the textbook) where we have a constant and an other explanatory variable X_i which takes the same value for all the observations in the sample (eg. $X_i = 2$). In this last case, there is exact multi-collinearity because the constant is equal to 0.5 times X_i. That is why we were unable to compute the OLS estimates.

Finally, note that even if we can not find the estimates in model (2) we can fit the model (3) if we omit the constant term:

(3) $S_i = \beta_2.MALE_i + \beta_3.FEMALE_i + u_i$

In the case of (3), we have again $FEMALE_i = 1 - MALE_i$, but the constant term ("1") is not included in the model. Hence there is no exact multi collinearity between the explanatory variables. However as the model (3) has no intercept the interpretation of the R^2 becomes complicated. Moreover, there is no reference category and you have to be careful when interpreting the point estimates for the two dummy variables.