ECON 6010 Statistical Methods

Dr. Fadi Fawaz
Department of Economics and Finance
Tennessee State University©

How to Work with Dummy Independent Variables

Questions:

1) How to include dummy variables in a regression?

2) How to interpret a coefficient on a dummy variable?

3) How to test hypotheses with dummy variables and interaction terms?

4) How to create a double-log functional form with dummy variables?

5) How to interpret a coefficient on a dummy variable with a log dependent variable?

1) How to include dummy variables in a regression?

Example:
You want to include Region of the United States in your earnings function regression. You obtain the variable GMREG from the Current Population Survey (CPS), and it has four possible values that the codebook maps to a region like this:

GMREG	Region
1	Northeast
2	Midwest
3	South
4	West

The data are sitting in an Excel file column like this:

GMREG
3
3
1
4
2
4
3

Obviously, you CANNOT use GMREG directly in a regression.

To incorporate region as a dummy variable, follow these steps:

1) Create Number of Categories – 1 new variables.

2) Include the Number of Categories – 1 variables in the regression.

The choice of which category to leave out (in this example, West) is *totally arbitrary* and has no effect on the final results. The actual coefficients of the regression equation do, of course, depend on the category left out (called the base case), but because you interpret a dummy variable coefficient relative to the base case, the predicted values end up the same.

2) How to interpret a coefficient on a dummy variable?

For a single dummy variable without an interaction term, the value of the coefficient tells you the change in the value of the dependent variable compared with the base case.

Example:

$$\text{Predicted Wage} = 10.2 + 1.3\text{Northeast} - 0.9\text{Midwest} - 1.7\text{South}$$

Consider cases:

- Person is from the Midwest. Then this person is predicted to make 10.2 – 0.9 = 9.3 (dollars per hour)—since the values for Northeast and South are zero.

- Southerners make $1.70 per hour less than Westerners.

Example:

$$\text{Predicted Wage} = 12.7 + 2.1\text{Education} - 1.8\text{Female}$$

- Holding Education constant, Female make $1.80 per hour less than Males (the base case).

If you have an interaction term, the situation get more complicated.

Example from Female.xls:

$$\text{Predicted Wage} = 0.076 + 0.88\text{Education} - 4.28\text{Female} + 0.16\text{Female*Education}$$

The Female coefficient is an intercept shifter and the interaction term affects the slope on Education.

A very useful approach to interpretation is to create a table and compute the predicted *Y* for given values of the included *X*'s.

NOTE: Sometimes, when you add an interaction term, the coefficient on the dummy variable becomes counter intuitively signed. For example, adding Black*Education might make the Black dummy variable coefficient positive. But the negative sign on the Black*Education must also be included when figuring out the effect of being black on predicted wage, and you will then obtain the expected result that blacks make less than non-blacks.

3) How to test hypotheses with dummy variables and interaction terms?

The *F*-test is the way to do this.

Obtain the SSR for the restricted and unrestricted models, compute the *F*-statistic (properly adjusting for the degrees of freedom in numerator and denominator), then find the *P*-value.

Example:

In a study of the determinants of wages, we want to see whether being female matters after controlling for education. We think that being female affects both the slope and intercept of the relationship between wages and education. Here are results from the unrestricted regression:

Predicted Wage = 0.076 + 0.88Education – 4.28Female + 0.16Female*Education
SSR = 419141
n = 8546

The corresponding restricted regression results are:

Predicted Wage = -1.53 + 0.92 Education
SSR = 430216.
n = 8546

For a test that there is no difference between male and female wages after controlling for education, the null hypothesis says that the coefficients on Female AND Female*Education together equal zero.

The test statistic is

$$F-\text{stat} = \frac{(430216 - 41941)/2}{419141/8542}$$
$$= \frac{5441}{49}$$
$$= 111.$$

The *P*-value is very small. We would reject the null hypothesis that being female doesn't matter for wage determination, after controlling for education.

4) How to create a double-log functional form with dummy variables?

You cannot take the natural log of a dummy variable because ln(0) is undefined.

Thus, you cannot create a completely double-log specification when you have dummy independent variables.

What is usually done is to take the natural log of the *Y* and continuous *X* variables, leaving the dummy variables untransformed.

Example (based on Duquette (1999)):

Own Units (or Levels) Specification:

$$\text{CharitableContributions}_i = \beta_0 + \beta_1 \text{Price}_i + \beta_2 \text{Disposable Income}_i + \beta_3 \text{Married}_i + \varepsilon_t$$

All variables are continuous except Married. Price measures the after-tax cost of a contribution. For example if for every dollar I give, my tax bill falls by 25 cents, the price is 0.75.

To make this a log-transformed model we take the natural logs of the continuous variables and leave the dummy variables alone.

Log Specification:

$$\ln \text{CharitableContributions}_i = \beta_0 + \beta_1 \ln \text{Price}_i + \beta_2 \ln \text{Disposable Income}_i + \beta_3 \text{Married}_i + \varepsilon_t$$

The b_1 and b_2 estimates could be interpreted as price and income elasticities respectively. The estimate of b_3 would be roughly the approximate percentage change in charitable contirbuitons for married tax payers versus unmarried tax payers. See the answer to the next question to learn how to precisely interpret the coefficient estimate for Married.

5) How to interpret a coefficient on a dummy variable with a log dependent variable?

The coefficient on a dummy variable with a log-transformed *Y* variable is interpreted as the percentage change in *Y* associated with having the dummy variable characteristic relative to the omitted category, with all other included *X* variables held fixed.

Example (Duquette (1999, p. 201)):
Predicted ln Charitable Giving =–4.46 - 1.3 lnPrice + 0.91 lnIncome +0.46 Married

Approximate Interpretation:
Predicted Charitable Giving is approximately 46 percent higher in for married tax payers, holding constant price of giving and income.

Exact Interpretation:

Actually, it is possible to obtain a more exact interpretation, and because the coefficient on the dummy variable is large, it matters. What we want is the percentage change in Charitable Giving between married and single taxpayers. This may be done as follows. Let PCG denote the predicted value of CharitableGiving (in the levels, not in natural logarithms) under a particular set of circumstances. We may then compute the percentage change in the predicted values between Married = 1 and Married = 0, with all other included Xs held constant, as

$$\frac{\left(PCG_{Married=1} - PCG_{Married=0}\right)}{PCG_{Married=0}}$$

We can express the percentage difference in predicted values as

$$\frac{PCG_{Married=1}}{PCG_{Married=0}} - 1$$

Now what we actually obtain from the log-specified regression equation is:

$$\begin{aligned}
&\ln(PCG_{Married=1}) - \ln(PCG_{Married=0}) = \\
&\ln\left(PCG_{Married=1} / PCG_{Married=1}\right) \\
&= b_3 \\
&= 0.46.
\end{aligned}$$

By the properties of logs, it then follows we can take the anti-log to obtain

$$\frac{PCG_{Married=1}}{PCG_{Married=1}} = \exp\left(b_3\right) = 1.58.$$

Then the exact computation of the percentage change is

$$\frac{\left(PCG_{Married=1} - PCG_{Married=0}\right)}{PCG_{Married=0}} = \frac{PCG_{Married=1}}{PCG_{Married=0}} - 1 = \exp(b_3) - 1 = \exp(0.46) - 1 = 1.58 - 1 = 0.58$$

Or, 58 percent. That is, married taxpayers, on average, make charitable contributions 58% percent higher than unmarried studies holding income and price constant.

The smaller the coefficient value, the closer the approximation will be to the exact computation because x gets closer to exp(x) –1 the smaller the x.

Reference:

Duquette, Christopher M. (1999) "Is Charitable Givign by Nonitemizers Responsive to Tax Incentives? New Evidence." *National Tax Journal* 52(2): 195-206.