ECON 5350 Class Notes
Dummy Variables and Functional Forms

1 Introduction

Although OLS is considered a linear estimator, it does not mean that the relationship between Y and X needs to be linear. In this section, we introduce several types of nonlinear functional forms, including those with dummy variables.

2 Dummy Variables

2.1 Comparing Two Means

Consider the regression model

$$y_i = \alpha + x_i'\beta + \delta d_i + \epsilon_i$$

for $i = 1, ..., n$ where d_i is a dummy or binary variable equal to one if the condition is satisfied and zero otherwise. The dummy variable can be used to contrast two means

$$\begin{aligned} E[y_i|x_i', d_i &= 0] = \alpha + x_i'\beta \\ E[y_i|x_i', d_i &= 1] = \alpha + x_i'\beta + \delta. \end{aligned}$$

A simple t test can then be used to test whether the conditional mean of y is different when $d = 0$ as opposed to $d = 1$. The null hypothesis for this test would be H_0: $\delta = 0$. An example will be presented below.

2.2 Several Categories

Some qualitative variables are naturally grouped into discrete categories (e.g., seasons of the year, religious affiliation, race, etc.). It also often useful to artificially categorize quantitative variables (e.g., income levels, education, age, etc.). Consider the regression model

$$y_i = \alpha + x_i'\beta + \delta_1 D_{1i} + \delta_2 D_{2i} + \delta_3 D_{3i} + \epsilon_i$$

with a set of four binary variables (D_{1i}, D_{2i}, D_{3i} and D_{4i}). Notice that one of the variables (D_{4i}) is excluded from the model to avoid the dummy-variable trap. Inclusion of D_{4i} would violate the Classical assumption that the X matrix is of full rank. This occurs because the four dummy variables will sum to a column of ones, which is perfectly collinear with the constant. By excluding D_{4i}, the δ coefficients are interpreted as

the change in y when moving from the j^{th} category ($D_{ji} = 1$) as compared to 4^{th} category ($D_{4i} = 1$).

2.2.1 Example. Testing for seasonality.

Consider two different models that test for seasonality in Y:

$$y_i = \alpha + x_i'\beta + \delta_1 D_{1i} + \delta_2 D_{2i} + \delta_3 D_{3i} + \epsilon_i \quad (1)$$

$$y_i = x_i'\beta + \theta_1 D_{1i} + \theta_2 D_{2i} + \theta_3 D_{3i} + \theta_4 D_{4i} + \epsilon_i. \quad (2)$$

Each equation avoids the dummy-variable trap – the first by excluding D_{4i} and the second by excluding the constant term α. Both equations generate the same goodness of fit, but the coefficients are interpreted differently. For equation (1) the test for seasonality is

$$H_0 : \delta_1 = \delta_2 = \delta_3 = 0$$

and for equation (2) the test is

$$H_0 : \theta_1 = \theta_2 = \theta_3 = \theta_4.$$

Both of these tests can be performed using the general F test developed in chapter 6 (provided the errors are normally distributed). It is instructive to relate the coefficients to one another

Category	Relationship between coefficients
$D_1 = 1$	$\theta_1 = \alpha + \delta_1 \Rightarrow \delta_1 = \theta_1 - \theta_4$
$D_2 = 1$	$\theta_2 = \alpha + \delta_2 \Rightarrow \delta_2 = \theta_2 - \theta_4$
$D_3 = 1$	$\theta_3 = \alpha + \delta_3 \Rightarrow \delta_3 = \theta_3 - \theta_4$
$D_4 = 1$	$\theta_4 = \alpha$

This highlights the fact that the δ coefficients are measured as the effect on Y, relative to the excluded category.

2.3 Interactive Effects

The dummy variables introduced in the previous two sections are intercept dummies because they cause parallel shifts in the regression line. Often, we want the slope of the regression line to change with some qualitative variable. For example, one might think that each extra inch of height for women is associated with a different change in weight than it is for men (i.e., the slope of the regression of weight on height is different for men than women). This can be incorporated using slope dummies or interaction terms.

Consider the following regression model

$$y_i = \alpha + \beta_1 x_i + \delta_1 D_i + \beta_2 (x_i D_i) + \epsilon_i. \tag{3}$$

The slope of the regression line when $D = 0$ equals β_1. The slope of the regression line when $D = 1$ equals $\beta_1 + \beta_2$. Therefore, to test whether the slope of the regression line is different when $D = 0$ versus $D = 1$ can be carried out by a simple t test of the null hypothesis H_0: $\beta_2 = 0$. The intercept dummy term $(\delta_1 D_i)$ is included so the two regression lines are free to have different intercepts.

2.4 Spline Regressions

Sometimes it is useful to let the slope of a regression line vary with some threshold value of a continuous explanatory variable. This so-called spline regression is basically a regression line with a kink(s). Consider the regression

$$y_i = \alpha + \beta x_i + \epsilon_i$$

with the additional restriction that the regression line is continuous everywhere but has a kink at x^0. The point x^0 is referred to as a knot. Note that the regression line is not differentiable at x^0. We need to introduce a dummy variable

$$D_i = \begin{array}{l} 1 \text{ if } x_i > x^0 \\ 0 \text{ otherwise} \end{array}.$$

The new spline regression model is

$$y_i = \alpha + \beta_1 x_i + \beta_2 (x_i D_i) + \beta_3 D_i + \epsilon_i \tag{4}$$

with the additional restriction that the line is continuous at $x_i = x^0$ (i.e., $\alpha + \beta_1 x^0 = \alpha + \beta_1 x^0 + \beta_2 x^0 + \beta_3$). Substituting the restriction $\beta_3 = -\beta_2 x^0$ back into equation (4) and rearranging gives

$$y_i = \alpha + \beta_1 x_i + \beta_2 D_i (x_i - x^0) + \epsilon_i.$$

The slope of the regression line when $x_i < x^0$ equals β_1. The slope of the regression line when $x_i > x^0$ equals $\beta_1 + \beta_2$.

2.5 MATLAB Example. Earnings Equation.

Consider the following (log) earnings equation

$$\ln(wage_i) = \beta_1 + \beta_2 Age_i + \beta_3 Age_i^2 + \beta_4 Grade_i + \beta_5 Married_i + \beta_6 (Married_i * Grade_i) + \epsilon_i$$

and a sample of 1000 men taken from the 1988 Current Population Survey. The code shows how to estimate (and graph) a regression model with both an intercept and a slope dummy variable. The code also estimates (and graphs) a spline regression with knots at $Grade_i = 12$ and $Grade_i = 16$. Finally, it tests to see if the slopes are equal across the different education levels (see MATLAB example 12).

3 Nonlinear Functional Forms

Below are some common nonlinear functional forms for regression models. Note that although some regression models may appear to violate Classical assumption #1 (i.e., the model is linear in the coefficients and error term), it is sometimes possible to take a linearizing transformation of the model (e.g., see the double-log form below).

3.1 Double-Log

Consider the following regression model

$$y_i = \beta_1 \prod_{j=2}^{k} x_{j,i}^{\beta_j} \exp(\epsilon_i)$$

which initially appears to be nonlinear in the parameters and the error term. However, by taking natural logs of both sides, we get

$$\ln(y_i) = \beta_1 + \beta_2 \ln(x_{2,i}) + \beta_3 \ln(x_{3,i}) + ... + \beta_k \ln(x_{k,i}) + \epsilon_i,$$

which is now obviously linear in β and ϵ. This is called the double-log functional form. The β coefficients can be interpreted as elasticities

$$\frac{\partial \ln(y_i)}{\partial \ln(x_{ki})} = \beta_k$$

or the percentage change in y_i for a one percent change in x_{ki}, all else equal.

3.2 Semi-Log

The semi-log functional form refers to either the dependent or some of the independent variables being logged. One example is

$$\ln(y_i) = \beta_1 + \beta_2 x_{2,i} + \beta_3 x_{3,i} + ... + \beta_k x_{k,i} + \epsilon_i.$$

3.2.1 Notes.

1. If, for instance, we introduced a time trend t as an explanatory variable, then $\partial \ln(y_i)/\partial t$ gives the conditional average growth rate of y_i.

2. β_2, for example, can be interpreted as the percentage change in y_i for a unit change in $x_{2,i}$.

3.3 Polynomial

An example of a polynomial functional form is

$$y_i = \beta_1 + \beta_2 x_{2,i} + \beta_3 x_{2,i}^2 + \beta_4 (x_{2,i} x_{3,i}) + \epsilon_i.$$

The coefficients do not measure the relevant partial derivatives. For example

$$\frac{\partial y_i}{\partial x_{2i}} = \beta_2 + 2\beta_3 x_{2,i} + \beta_4 x_{3,i}$$

which needs to be evaluated at some value for x_2 and x_3. High-order polynomial functional forms can provide excellent goodness of fit within the sample, however, the out-of-sample fit is often poor. Rarely does theory call for anything over second-order polynomials.