

POL 681 Lecture Notes: Regression with Categorical Covariates ... or how smart are your dummies?

The regression model we've considered to this point has primarily dealt with continuous or "quantitative" independent variables. Hence, in the model

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i},$$

if X_1 and X_2 are continuous, then the natural interpretation for a slope coefficient is forthcoming: for an incremental (or unit) change in X , the expected value of Y changes (increases or decreases) by about $\hat{\beta}$ amount. Further, as we interpret these coefficients as partial regression slopes, we can make statements like "holding one variable constant," the relationship between Y and the other X is given by $\hat{\beta}$. In any event, since X can range continuously (theoretically), we can use this information to trace out the relationship between X and Y . Graphically, this is what gives us our slope.

What about "qualitative" or nominal outcomes? These are the kinds of variables that take on two or more values but the scores of the variable are assumed to represent categories. "Dummy" variables are the most well known type of this kind of variable. Such a variable may represent gender, racial classifications, party classifications, or so on. Fortunately, such covariates pose few problems for the regression model. The main issues involve coding categorical variables and interpreting them.

The motivation for including dummy variables, as Fox notes, is the same as including any kind of covariate. Presumably, one may have some hypothesis about differences across levels of the categorical covariate or further, one may believe that accounting for the categorical covariate will help account for greater variation in the outcome variable. Either way, it may be natural to include categorical covariates into a regression model.

1 The Bottom-Line Interpretation

Let's first consider only dichotomous categorical variables; that is, variables having only two values. We'll call this variable D . Out of convenience, we code these outcomes as 0 and 1 (though it makes no difference what the coding scheme is—it's arbitrary). The key thing to note is that D only has two outcomes! That means that in terms of predicted values, changes to D can *only* result in a change from one value to another value. That is, there are only two possible outcomes for \hat{Y} .

Thus, in the model

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_{1i},$$

the predicted value of \hat{Y} when $D = 1$ is given by

$$\hat{\beta}_0 + \hat{\beta}_1 1.$$

When $D = 0$, the predicted value of Y is given by

$$\hat{\beta}_0 + \hat{\beta}_1 0,$$

or $\hat{\beta}_0$. Thus, in a bivariate model with one dummy variable, there will only be two predicted values! Note also that since our dummy variable has a natural 0 point (which we arbitrarily defined), the constant term has a natural interpretation: it gives us the predicted Y for the condition when $D = 0$.

This illustration is straightforward and illustrates a fundamental point: dummy variables, constructed in this way, *are nothing more than differences in intercepts*. In the previous model, the coefficient $\hat{\beta}_1$ tells us the expected increase (if the coefficient is positive) in Y for observations where $D = 1$. Note also that since the model above has no continuous (or quantitative) covariates, there are no slope coefficients. Thus, if you plotted \hat{Y} with respect to D , you would simply get “two dots.”

To illustrate ideas, consider the data in the companion Word tutorial named `dummyvariables.doc`. [Go to Word handout, which is stored on the website.]

2 More Motivation

Sometimes, we include categorical covariates to account for variation in Y that is a function of some group-specific or attribute-specific factor. Often, we can only measure these type of factors at a nominal level (gender, race, group affiliation, etc.). Other times, we may wish to “control” for these kinds of factors. For example, if we believe that gender has an independent “effect” in voting models above and beyond what, say, some measure of ideology has, then it may make sense to account for this attribute in our model. Since gender is usually recorded dichotomously, this would motivate the inclusion of a dummy variable recording the sex of the respondent (1 for females, 0 for males, or visa versa . . . it makes no difference).

Frequently, our concern with dummy variables will be in conjunction with a concern for quantitative variables. Our models will have a mixture of both kinds of variables. To explain, suppose we have a model with one quantitative variable and one dummy variable. This model is given by:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 D_{1i},$$

where $\hat{\beta}_1$ is the coefficient for our quantitative variable X and $\hat{\beta}_2$ is the coefficient for our dummy variable D . Unlike the model in the previous section which yielded only differences in intercepts, this model will give us the following: different intercepts *but* parallel slopes.

Suppose that $D = 1$, then the regression response function for this condition is:

$$\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 1.$$

Rearranging, we find the following:

$$(\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 X_{1i}.$$

In contrast, when $D = 0$, the regression response function is given by

$$\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 0,$$

which is equivalent to

$$(\hat{\beta}_0) + \hat{\beta}_1 X_{1i}.$$

It is clear, the “effect” of the dummy variable in the model is to act as a “contrast” or “offset” for the two groups. When $D = 1$, the coefficient $\hat{\beta}_2$ represents the increase in the y -intercept attributable to D ; when $D = 0$, the constant term (i.e. $\hat{\beta}_0$) gives us the intercept for this condition. The important thing to note, however, is the following: *the effect of X_1 on Y is the same for both groups*. Thus, the dummy variable only reflects a difference in intercepts while the slope coefficient will be equivalent for the two groups. The substantive interpretation? The variable X_1 has the same “effect” for both conditions of the dummy variable, it is just that the intercept differs for the two groups.

A model like this one is known as a “parallel” slopes regression. The reason is simple: plot the predicted regression function and you’ll get two functions—one for $D = 1$ and one for $D = 0$ —each having identical slopes.

To fix ideas, consider the data in the companion Word tutorial named `dummyvariables.doc`. [Go to Word handout, which is stored on the website.]

3 Polytomous Variables

The main issues of polytomous variables are covered in the tutorial.