Qualitative Variables and Regression Analysis

Allin Cottrell

October 3, 2011

1 Introduction

In the context of regression analysis we usually think of the variables are being quantitative—monetary magnitudes, years of experience, the percentage of people having some characteristic of interest, and so on. Sometimes, however, we want to bring qualitative variables into play. For example, after allowing for differences attributable to experience and education level, does gender, or marital status, make a difference to people's pay? Does race make a difference to pay, or to the chance of becoming unemployed? Did the coming of NAFTA make a significant difference to the trade patterns of the USA? In all of these cases the variable we're interested in is qualitative or categorical; it can be given a numerical coding of some sort but in itself it is non-numerical.

Such variables can be brought within the scope of regression analysis using the method of *dummy variables*. This method is quite general, but let's start with the simplest case, where the qualitative variable in question is a *binary variable*, having only two possible values (male versus female, pre-NAFTA versus post-NAFTA).

The standard approach is to code the binary variable with the values 0 and 1. For instance we might make a gender dummy variable with the value 1 for males in our sample and 0 for females, or make a NAFTA dummy variable by assigning a 0 in years prior to NAFTA and a 1 in years when NAFTA was in force.

2 Gender and salary

Consider the gender example. Suppose we have data on a sample of men and women, giving their years of work experience and their salaries. We'd expect salary to increase with experience, but we'd like to know whether, controlling for experience, gender makes any difference to pay. Let y_i denote individual i's salary and x_i denote his or her years of experience. Let D_i (our gender dummy) be 1 for all men in the sample and 0 for the women. (We could assign the 0s and 1s the other way round; it makes no substantive difference, we just have to remember which way round it is when we come to interpret the results.) Now we estimate (say, using OLS) the model

$$y_i = \alpha + \beta x_i + \gamma D_i + \epsilon_i \tag{1}$$

In effect, we're getting "two regressions for the price of one". Think about the men in the sample. Since they all have a value of 1 for D_i, equation (1) becomes

$$\begin{aligned} y_i &= \alpha + \beta x_i + \gamma \cdot 1 + \epsilon_i \\ &= \alpha + \beta x_i + \gamma + \epsilon_i \\ &= (\alpha + \gamma) + \beta x_i + \epsilon_i \end{aligned}$$

Since the women all have $D_i = 0$, their version of the equation is

$$\begin{aligned} y_i &= \alpha + \beta x_i + \gamma \cdot 0 + \epsilon_i \\ &= \alpha + \beta x_i + \epsilon_i \end{aligned}$$

Thus the male and female variants of our model have different intercepts, $\alpha + \gamma$ for the men and just α for the women.

Suppose we conjecture that men might be paid more, after allowing for experience. If this is true, we'd expect it to show up in the form of a positive value of our estimate for the parameter γ. We can test the idea that gender makes a difference by testing the null hypothesis $H_0 : \gamma = 0$. If our estimate of γ is positive and statistically significant we reject the null and conclude that men are paid more.

We could, of course, simply calculate the mean salary of the men in the sample and the mean for women and compare them (perhaps doing a t-test for the difference of two means). But that would *not* accomplish the same as the above approach, since it would not control for years of experience. It could be that male salaries are higher on average, but the men also have more experience on average, and the difference in salary by gender is entirely explained by difference in experience levels. By running a regression including both experience and a gender dummy variable we can distinguish this possibility from the possibility that, over and above any effects of differential experience levels, there is a systematic difference by gender.

Here's output from a regression of this sort run in `gretl`, using `data7-2` from among the Ramanathan practice files. Actually, rather than experience I'm using EDUC (years of education beyond 8th grade when hired) as the control variable. As you can see, in this instance men were paid more, controlling for education level. The GENDER coefficient is positive and significant; it appears that men were paid about $550 more than women with the same educational level.

OLS estimates using the 49 observations 1–49
Dependent variable: WAGE

Variable	Coefficient	Std. Error	t-statistic	p-value
const	856.231188	227.835435	3.7581	0.000481
EDUC	108.061579	32.439606	3.3312	0.001712
GENDER	549.072697	152.732420	3.5950	0.000788

Mean of dep. var.	1820.204082	S.D. of dep. variable	648.268719
ESS	13077037.992324	Std Err of Resid. ($\hat{\sigma}$)	533.182365
R^2	0.351727	$\bar{R}^2$	0.323541
F-statistic (2, 46)	12.478873	p-value for F()	0.000047

3 Extending the idea

There are two main ways in which the basic idea of dummy variables can be extended:

- Allowing for qualitative variables with more than two values.
- Allowing for difference in slope, as well as difference of intercept, across qualitative categories.

An example of the first sort of extension might be "race". Suppose we have information that places people in one of four categories, White, Black, Hispanic and Other, and we want to make use of this along with quantitative information in a regression analysis.

The rule is that to code n categories we need $n - 1$ dummy variables, so in this case we need three "race dummies". We have to choose one of the categories as the "control"; members of this group will be assigned a 0 on all the dummy variables. Beyond that, we need to arrange for each category to be given a unique pattern of 0s and 1s on the set of dummy variables. One way of doing this is shown in the following table, which defines the three variables $R1$, $R2$ and $R3$.

	$R1$	$R2$	$R3$
White	0	0	0
Black	1	0	0
Hispanic	0	1	0
Other	0	0	1

You might ask, Why do we need all those variables? Why can't we just define *one* race dummy, and assign (say) values of 0 for Whites, 1 for Blacks, 2 for Hispanics and 3 for Others? Unfortunately this will not do what we want. Consider a slightly simpler variant—a three-way comparison of Whites, Blacks and Hispanics, where we define one variable R with values of 0, 1 and 2 for Whites, Blacks and Hispanics respectively. Using the same reasoning as given above in relation to model (1) we'd have (for given quantitative variables x and y):

$$\begin{array}{ll}
\text{Overall:} & y_i = \alpha + \beta x_i + \gamma R_i + \epsilon_i \\[6pt]
\text{White:} & y_i = \alpha + \beta x_i + \gamma \cdot 0 + \epsilon_i \\
 & y_i = \alpha + \beta x_i + +\epsilon_i \\[6pt]
\text{Black:} & y_i = \alpha + \beta x_i + \gamma \cdot 1 + \epsilon_i \\
 & y_i = (\alpha + \gamma) + \beta x_i + +\epsilon_i \\[6pt]
\text{Hispanic:} & y_i = \alpha + \beta x_i + \gamma \cdot 2 + \epsilon_i \\
 & y_i = (\alpha + 2\gamma) + \beta x_i + +\epsilon_i
\end{array}$$

We're allowing for three different intercepts OK, but we're constraining the result: we're insisting that whatever the difference in intercept between Whites and Blacks (namely γ), the difference in intercept between Whites and Hispanics is exactly twice as big (2γ). But there's no reason to expect this pattern. In general, we want to allow the intercepts for the three (or more) groups to differ arbitrarily—and that requires the use of $n - 1$ dummy variables.

Let's see what happens if we define two dummies, $R1$ and $R2$, to cover the three "race" categories as shown below:

	$R1$	$R2$
White	0	0
Black	1	0
Hispanic	0	1

The general model is

$$y_i = \alpha + \beta x_i + \gamma R1_i + \delta R2_i + \epsilon_i$$

and it breaks out as follows for the three groups:

$$\begin{array}{ll}
\text{White:} & y_i = \alpha + \beta x_i + \gamma \cdot 0 + \delta \cdot 0 + \epsilon_i \\
 & y_i = \alpha + \beta x_i + \epsilon_i \\[6pt]
\text{Black:} & y_i = \alpha + \beta x_i + \gamma \cdot 1 + \delta \cdot 0 + \epsilon_i \\
 & y_i = (\alpha + \gamma) + \beta x_i + \epsilon_i \\[6pt]
\text{Hispanic:} & y_i = \alpha + \beta x_i + \gamma \cdot 0 + \delta \cdot 1 + \epsilon_i \\
 & y_i = (\alpha + \delta) + \beta x_i + \epsilon_i
\end{array}$$

Thus we have three independent intercepts, α, $\alpha + \gamma$, and $\alpha + \delta$. The null hypothesis "race makes no difference" translates to $H_0 : \gamma = \delta = 0$, which can be tested using an F-test.

Translating codings

This raises a practical issue. Suppose we have a qualitative variable that is coded as 0, 1, 2 and so on (as is the case with a lot of data available from government sources such as the Bureau of the Census). We saw above that we can't use such a coding as is, for the purposes of regression analysis; we'll have to convert the information into an appropriate set of 0/1 dummy variables first.

You could do this using formulas in a spreadsheet, but it's probably easier to do it in `gretl`. Suppose we have a variable in the current dataset called RACE, which is coded 0, 1, 2 and so on. We want to create a dummy called R1 which has value 1 for all cases where RACE equals 1, and 0 otherwise. Under the "Variable" menu, choose the item "Define new variable". A dialog box comes up where you enter the formula for the new variable. In this case we'd type `R1 = (RACE=1)`. The first "=" here is the equals of assignment; it is being used to define the new variable R1. The second "=" is being used as a Boolean (logical) operator. That is, the expression `(RACE=1)` gives a result of 1 when the condition evaluates as true, i.e. where RACE does equal 1, and 0 when the condition is false, i.e. for any other values of RACE.

Another example: Consider the categorization of educational attainment offered in the Current Population Survey.

```
00 .Children
31 .Less than 1st grade
32 .1st, 2nd, 3rd, or 4th grade
33 .5th or 6th grade
34 .7th and 8th grade
35 .9th grade
36 .10th grade
37 .11th grade
38 .12th grade no diploma
39 .High school graduate
40 .Some college but no degree
41 .Associates degree-occup./vocational
42 .Associates degree-academic program
43 .Bachelors degree(BA,AB,BS)
44 .Masters degree(MA,MS,MEng,MEd,MSW,MBA)
45 .Prof. school degree (MD,DDS,DVM,LLB,JD)
46 .Doctorate degree(PhD,EdD)
```

Suppose we want to make out of this a three-way classification, the categories being "no High school diploma", "High school diploma but no Bachelors Degree", and "Bachelors degree or higher". If the variable shown above is called AHGA, then in `gretl` we could define two dummy variables thus:

```
E1 = (AHGA>38) & (AHGA<43)
E2 = AHGA > 42
```

The "&" (logical AND) in the first formula means that E1 will get value 1 only if both conditions, `(AHGA>38)` and `(AHGA<43)`, are satisfied, corresponding to "High school diploma but no Bachelors Degree", while the definition of E2 corresponds to "Bachelors degree or higher". Those without a High school diploma are the control group, with 0s for both E1 and E2.

4 Allowing for differing slopes

The regression models above allow the intercept of the regression to differ across qualitative categories. In all cases so far, however, we have imposed a common slope, β, with respect to the (quantitative) independent variable x. We might want to allow the slope to differ too. For example, it might be that while men and women are both paid more highly if they have more experience or education, the *degree* to which experience or education brings higher pay may differ for men and women. Note that this is a different point from simply saying that men and women at the same level of education or experience are paid differently.

To allow for this sort of thing we can define an *interaction term*, by multiplying a dummy variable into x. Let's go back to equation (1) but add a new variable S such that $S_i = D_i x_i$. The model then becomes

$$y_i = \alpha + \beta x_i + \gamma D_i + \delta S_i + \epsilon_i \tag{2}$$

which breaks out for men and women as:

$$\begin{aligned}
\text{Men:}\quad & y_i = \alpha + \beta x_i + \gamma \cdot 1 + \delta x_i \cdot 1 + \epsilon_i \\
& y_i = \alpha + \beta x_i + \gamma + \delta x_i + \epsilon_i \\
& y_i = (\alpha + \gamma) + (\beta + \delta) x_i + \epsilon_i \\
\text{Women:}\quad & y_i = \alpha + \beta x_i + \gamma \cdot 0 + \delta x_i \cdot 0 + \epsilon_i \\
& y_i = \alpha + \beta x_i + \epsilon_i
\end{aligned}$$

This now allows for different slopes ($\beta + \delta$ for men, just β for women) as well as different intercepts. To test whether gender makes any difference (either to the intercept or the slope) we would use an F-test on $H_0 : \gamma = \delta = 0$.