Dummy Explanatory Variables

Christopher Taber

Department of Economics
University of Wisconsin-Madison

April 5, 2010

Categorical Variables

Lets go back to something we thought about very early on in this course: the difference in average wages between men and women

Suppose you want to test whether men make more money than women

That is you have the following null hypothesis

$$H_0 : E(W \mid \text{Male}) = E(W \mid \text{Female})$$

where W is hourly earnings.

How do you do this?

You could take means for each group and calculate

$$\frac{\bar{W}_m - \bar{W}_f}{\sqrt{se\left(\bar{W}_m\right)^2 + se\left(\bar{W}_f\right)^2}}$$

It turns out that there is an easier way.

Suppose we have data on men's and women's wages.

We want to run a regression, but how do we do that?

“Man” and “Woman” are categories, not numbers

To run a regression we need a number

Solution: Turn it into a number as a **Dummy Variable**

Define

$$m_i \quad = \quad \begin{cases} 1 & \text{Person is male} \\ 0 & \text{Person is female.} \end{cases}$$

Now let's see if regression analysis can be useful.

We will think of this in a "descriptive" way

Let

$$E\left(W_i \mid m_i\right) = \beta_0 + \beta_1 m_i.$$

Why is this useful?

Now notice that

$$\begin{aligned} E\left(W_i \mid \text{Male}\right) &= E\left(W_i \mid m_i = 1\right) \\ &= \beta_0 + \beta_1 \\ E\left(W_i \mid \text{Female}\right) &= E\left(W_i \mid m_i = 0\right) \\ &= \beta_0 \end{aligned}$$

Solving out this means that

$$\begin{aligned} \beta_0 &= E\left(W_i \mid \text{Female}\right) \\ \beta_1 &= E\left(W_i \mid \text{Male}\right) - E\left(W_i \mid \text{Female}\right) \end{aligned}$$

But then testing $H_0 : E\,(W \mid \text{Male}) = E\,(W \mid \text{female})$ is equivalent to testing

$$H_0 : \beta_1 = 0$$

We already know how to do this.

Lets see.

Adding Conditioning Variables

But that isn't all.

We might be worried that women have less labor market experience than men.

An interesting null hypothesis might be

$$H_0 : E\left(W \mid \text{Male,Experience}\right) = E\left(W \mid \text{Female,Experience}\right)$$

That is, comparing men and women with the same level of experience, do they earn the same amount of money?

This is easy to do, we just write the model as

$$E\left(W_i \mid m_i\right) = \beta_0 + \beta_1 m_i + \beta_2 Exp_i + \beta_3 Exp_i^2.$$

Then testing whether $\beta_1 = 0$ tests exactly what we want.

Why stop there?

We can condition on whatever we want

$$E\left(W_i \mid m_i\right) = \beta_0 + \beta_1 m_i + \beta_2 X_{i2} + ... + \beta_K X_{iK}.$$

Lets look at some examples.

Interactions

Back to the male/female example

The regression model

$$E\left(W_i \mid m_i\right) = \beta_0 + \beta_1 m_i + \beta_2 Exp_i.$$

imposes that the “return to experience” is the same for women as it is for men.

You might think that this is not quite right.

For example if you are worried about glass ceilings then you might think the slope is different for men than it is for women.

One could think of just running separate regressions for men and for women

$$\begin{aligned} E\left(W \mid \text{Male,Experience}\right) &= \beta_0^m + \beta_1^m Exp_i \\ E\left(W \mid \text{Female,Experience}\right) &= \beta_0^f + \beta_1^f Exp_i \end{aligned}$$

Lets see what this looks like

But what if I want to test whether these things are the same?

That is I might want to run a joint test of whether men and women face the same earnings profile

The key here is an interaction.

Think about the model

$$E\left(W_i \mid \text{Gender,Experience}\right) = \beta_0 + \beta_1 m_i + \beta_2 Exp_i + \beta_3 \left(m_i \times Exp_i\right)$$

Then

$$\begin{aligned}
E\left(W \mid \text{Male,Experience}\right) &= \beta_0^m + \beta_1^m Exp_i \\
&= E\left(W_i \mid m_i = 1, Exp_i\right) \\
&= \beta_0 + \beta_1 + \beta_2 Exp_i + \beta_3 Exp_i \\
E\left(W \mid \text{Female,Experience}\right) &= \beta_0^f + \beta_1^f Exp_i \\
&= E\left(W_i \mid m_i = 0, Exp_i\right) \\
&= \beta_0 + \beta_2 Exp_i
\end{aligned}$$

Thus

$$\begin{aligned} \beta_0 &= \beta_0^f \\ \beta_1 &= \beta_0^m - \beta_0^f \\ \beta_2 &= \beta_2^f \\ \beta_3 &= \beta_2^m - \beta_2^f \end{aligned}$$

Testing that the profiles are the same is equivalent to testing the joint null hypothesis:

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ \beta_3 &= 0 \end{aligned}$$

The Dummy Variable Trap

Think about the following exercise

What would happen if we constructed the new dummy variable

$$f_i = \begin{cases} 1 & \text{Person i is female} \\ 0 & \text{Person i is male} \end{cases}$$

What if we then tried to run a regression based on

$$E\left(W_i \mid Gender\right) = \beta_0 + \beta_1 m_i + \beta_2 f_i?$$

It turns out that this will not work.

We have perfect multicollinearity because

$$m_i = 1 - f_i$$

This is called the **Dummy Variable Trap**

This makes sense if you think about it.

There are really only two pieces of information in the population

$$\begin{aligned} E\left(W_i \mid \text{Male}\right) &= \beta_0 + \beta_1 \\ E\left(W_i \mid \text{Female}\right) &= \beta_0 + \beta_2 \end{aligned}$$

We have 3 parameters and 2 equations

Clearly the model is not identified so it makes sense that you would have problems

Stata is smart about this kind of thing though

More than Two Categories

So far we have dealt with categorical variables with only 2 categories, but this is clearly not the only interesting case

For example think about race where we could think of (at least) 5 groups

Race could be

African American
Asian
Hispanic
Native American
All others

We are still going to have the dummy variable trap, but in this case it means we must omit 1 category

Let B, A_i, H_i, N_i be dummy variables for black, asian, hispanic, and native american respectively.

That is for example

$$B_i = \begin{cases} 1 & \text{Person i is African American} \\ 0 & \text{otherwise} \end{cases}$$

Then we can think of the regression

$$E\left(W_i \mid \text{Race}\right) = \beta_0 + \beta_1 B_i + \beta_2 A_i + \beta_3 H_i + \beta_4 N_i$$

Note that we have 5 basic population equations (for the 5 races) and 5 parameters so we seemed to have solved the dummy variable trap problem.

How do we interpret the parameters?

$$
\begin{aligned}
E\left(W_i \mid \text{African American}\right) &= \beta_0 + \beta_1 \\
E\left(W_i \mid \text{Asian}\right) &= \beta_0 + \beta_2 \\
E\left(W_i \mid \text{Hispanic}\right) &= \beta_0 + \beta_3 \\
E\left(W_i \mid \text{Native American}\right) &= \beta_0 + \beta_4 \\
E\left(W_i \mid \text{All Others}\right) &= \beta_0
\end{aligned}
$$

Thus solving out one can show that

$$
\begin{aligned}
\beta_0 &= E\left(W_i \mid \text{All Others}\right) \\
\beta_1 &= E\left(W_i \mid \text{African American}\right) - E\left(W_i \mid \text{All Others}\right) \\
\beta_2 &= E\left(W_i \mid \text{Asian}\right) - E\left(W_i \mid \text{All Others}\right) \\
\beta_3 &= E\left(W_i \mid \text{Hispanic}\right) - E\left(W_i \mid \text{All Others}\right) \\
\beta_4 &= E\left(W_i \mid \text{Native American}\right) - E\left(W_i \mid \text{All Others}\right)
\end{aligned}
$$

Thus the left out group matters a lot in the interpretation of the parameters

Note that

$$\beta_1 - \beta_3 = E\left(W_i \mid \text{African American}\right) - E\left(W_i \mid \text{Hispanic}\right)$$

so if we want to test whether Hispanics and Blacks earn the same, this is easy to do

Lets look at some examples.

Multiple Categorical Variables

Now what about more than one categorical variable at a time.

For example what about race and gender?

Lets just focus on the african american gap by putting all other groups together.

What would happen if we just thought about the model as:

$$E\left(W_i \mid \text{Race,Gender}\right) = \beta_0 + \beta_1 m_i + \beta_2 B_i$$

Note that in this model

$$\begin{aligned} E\left(W_i \mid \text{White Male}\right) &= \beta_0 + \beta_1 \\ E\left(W_i \mid \text{White Female}\right) &= \beta_0 \\ E\left(W_i \mid \text{Black Male}\right) &= \beta_0 + \beta_1 + \beta_2 \\ E\left(W_i \mid \text{Black Female}\right) &= \beta_0 + \beta_2 \end{aligned}$$

Now actually we have 4 equations and three parameters so we can’t solve out exactly.

Note that

$$E\left(W_i \mid \text{White Male}\right) - E\left(W_i \mid \text{White Female}\right) = \beta_1$$

and

$$E\left(W_i \mid \text{Black Male}\right) - E\left(W_i \mid \text{Black Female}\right) = \beta_1$$

We have imposed this, but it may not be true.

How do we relax this?

An interaction between B_i and m_i.

$$E\left(W_i \mid \text{Race,Gender}\right) = \beta_0 + \beta_1 m_i + \beta_2 B_i + \beta_3\left(m_i \times B_i\right)$$

then

$$\begin{aligned} E\left(W_i \mid \text{White Male}\right) &= \beta_0 + \beta_1 \\ E\left(W_i \mid \text{White Female}\right) &= \beta_0 \\ E\left(W_i \mid \text{Black Male}\right) &= \beta_0 + \beta_1 + \beta_2 + \beta_3 \\ E\left(W_i \mid \text{Black Female}\right) &= \beta_0 + \beta_2 \end{aligned}$$

Now

$$E\left(W_i \mid \text{White Male}\right) - E\left(W_i \mid \text{White Female}\right) = \beta_1$$

and

$$E\left(W_i \mid \text{Black Male}\right) - E\left(W_i \mid \text{Black Female}\right) = \beta_1 + \beta_3$$