Dummy variables in multiple variable regression model

1. Additive dummy variables

In the previous handout we considered the following regression model:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

and we interpreted the coefficients by partially differentiating the dependent variable with respect to each explanatory variable

$$\frac{\partial y_i}{\partial x_{1i}} = \beta_1 = \frac{\text{Change in } y_i}{\text{Unit increase in } x_1}\Bigg|_{\text{Holding all else constant}}$$

However, in many cases we do not have actual data for many variables that we want to include into a regression model. For example, in an equation for wage we may want to allow for the possibility that females appeared to be paid less money for the same job. In which case we may want to specify a model:

$$\ln(w_i) = \alpha + \beta_1 School_i + \beta_2 Female_i + \varepsilon_i \qquad (1)$$

where, *School* = Number of years of schooling, and $Female = 1$ if female, 0 if male.

The variable *Female* is known as an additive dummy variable and has the effect of vertically shifting the regression line. In all models with dummy variables the best way to proceed is write out the model for each of the categories to which the dummy variable relates. So in our case the categorical variable would be gender (which has two categories Males and Females). For Males (when *Females*=0), we have from (1):

$$\ln(w_i) = \alpha + \beta_1 School_i + \varepsilon_i \qquad (2a)$$

Whereas for females (when *Females*=1), we have from (1)

$$\ln(w_i) = \alpha + \beta_1 School_i + \beta_2 + \varepsilon_i$$

that is,

$$\ln(w_i) = (\alpha + \beta_2) + \beta_1 School_i + \varepsilon_i \qquad (2b)$$

and so the intercept is α in equation (2a) and is $(\alpha + \beta_2)$ in equation (2b). So the two lines are parallel (as they have the same slope coefficients), but they cross the vertical axis at different points. If we take expected values of equation (2a) we get:

$$E\left[\ln(w^M)\right] = \alpha + \beta_1 School_i$$

and taking expected values of equation (2b) we get

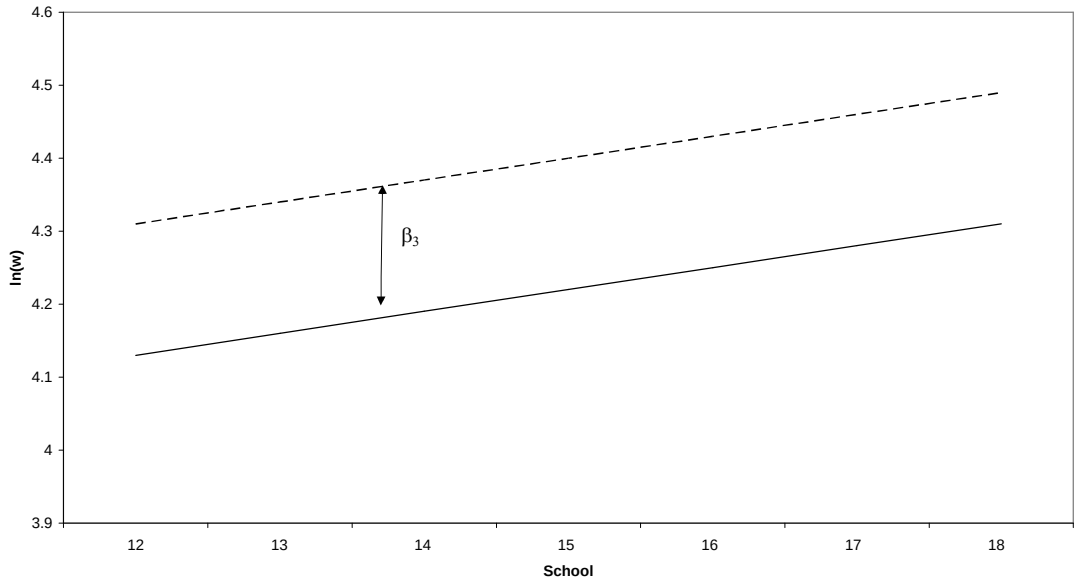
$$E\left[\ln(w)^F\right] = (\alpha + \beta_2) + \beta_1 School_i$$

Now the difference between these two expressions for the same value of *School* is:

$$E\left[\ln(w)^F\right] - E\left[\ln(w)^M\right] = \beta_2$$

so β_2 is of the change in the expected value of ln(w_i) for a female compared to a male for a given value of *School*. Pictorially this is shown in Figure 1, where the dashed line represents the regression line for males and the solid line that for females.

Figure 1: The role of an additive dummy variable



If instead of estimating equation (1), we estimated

$$\ln(w_i) = \delta_0 + \delta_1 School_i + \delta_2 Male_i + \varepsilon_i \qquad (3)$$

where $Male = 1$ if male, 0 if female , as $Male = 1 - Female$ then

$$\ln(w_i) = \delta_0 + \delta_1 School_i + \delta_2 (1 - Female_i) + \varepsilon_i$$

$$\ln(w_i) = \delta_0 + \delta_1 School_i + \delta_2 - \delta_2 Female_i + \varepsilon_i$$

$$\ln(w_i) = (\delta_0 + \delta_2) + \delta_1 School_i - \delta_2 Female_i + \varepsilon_i \qquad (4)$$

Comparing equation (1) with equation (4) as the variables on the left and right hand side of the equation are identical the coefficients must be identical, in which case:

$$\delta_2 = -\beta_2,\ \delta_1 = \beta_1, \delta_0 = \alpha - \delta_2$$

and so by using *Male* in place of *Female* in equation (1), all that happens is the coefficient on the *Male* dummy will be minus that on the *Female* dummy and the coefficient on the intercept will shift accordingly to allow for the fact that a Female is now the default case.

For the categorical variable gender (in which there are two categories), you include 1 dummy variable in the regression model (in our example *Female*), so that Male is the

default and the coefficient on Female is the change in the expected value of the dependent variable (for given values of the other variables) for females relative to males.

In general, if there is a categorical variable with *s* categories, then you include *s*-1 dummy variables and the omitted category is the default and the coefficient on any included category is then measured relative to the default (omitted) category.

2. Multiplicative dummy variables

In the equation for wage we may also want to allow for the possibility that female pay, relative to male pay, varies with the level of schooling. In which case we may want to specify a model:

$$\ln(w_i) = \alpha + \beta_1 School_i + \beta_2 Female_i + \beta_3 (Female_i \times School_i) + \varepsilon_i \qquad (5)$$

where *School* = Number of years of schooling and

$Female = 1$ if person is female, 0 if male . The variable *Female* is known as an additive dummy variable and $(Female_i \times School_i)$ is known as a multiplicative dummy variable. The multiplicative dummy variable has the effect of rotating the regression line, so for Males (when *Females*=0) and we have from (5):

$$\ln(w_i) = \alpha + \beta_1 School_i + \varepsilon_i \qquad (6)$$

Whereas for females (when *Females*=1) and we have from (5)

$$\ln(w_i) = \alpha + \beta_1 School_i + \beta_2 + \beta_3 School_i + \varepsilon_i$$

that is,

$$\ln(w_i) = (\alpha + \beta_2) + (\beta_1 + \beta_3) School_i + \varepsilon_i \qquad (7)$$

and so the intercept is α and the coefficient on *School* is β_1 in equation (6) and the intercept is $(\alpha + \beta_2)$ and the coefficient on *School* is $(\beta_1 + \beta_3)$ in equation (7). So β_2 is the change in the expected value of ln(w_i) for a female compared to a male for *School*=0, while,

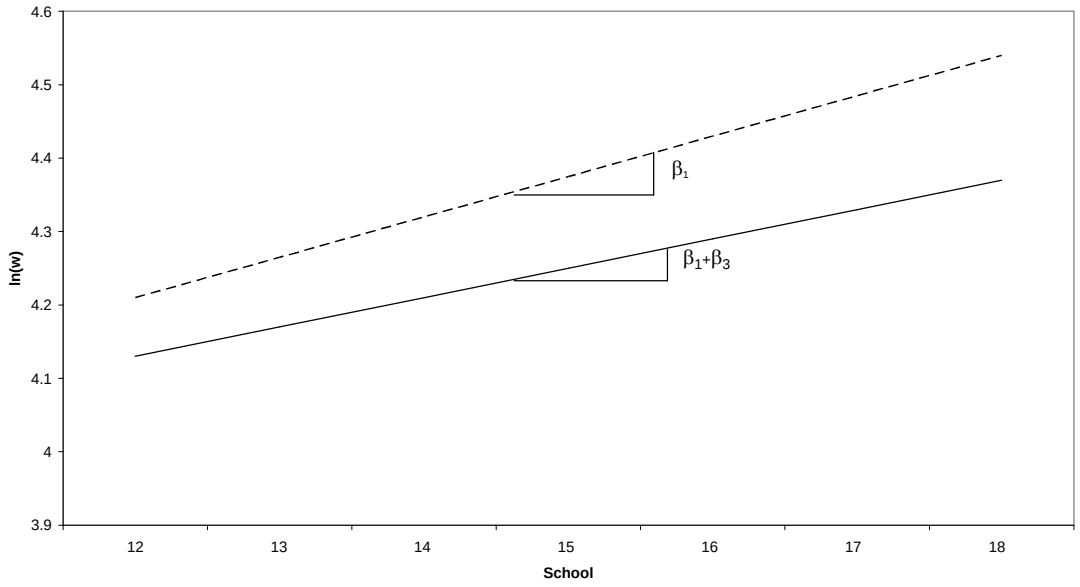
$$\frac{\partial \ln(w_i)}{\partial School} = \beta_1 \Big|_{\text{All else constant and Male}}$$

$$\frac{\partial \ln(w_i)}{\partial School} = (\beta_1 + \beta_3) \Big|_{\text{All else constant and Female}}$$

and β_3 is the increase in the semi-elasticity of wages with respect to an extra year of schooling for females compared to males. Pictorially this is shown in Figure 2, where

the dashed line represents the regression line for males and the solid line that for females.

Figure 2: The role of an multiplicative dummy variable



As gender has two categories, you can only include 1 multiplicative dummy variable in our regression model ($Female \times School$), so that Male is the default and the coefficient on $Female \times School$ is the change in the coefficient on *School* (for given values of the other variables) for females compared to males.

In general, if there is a categorical variable with *s* categories, then you include *s*-1 multiplicative dummy variables (multiplied by *School*) and the omitted category is the default and the coefficient on any of the multiplicative dummy variables is measured relative to the default (omitted) category.

3. Interactive dummy variables

In the equation for wage we may also want to allow for the possibility that two dummy variables could be multiplied together, for example, to allow in the model females to be paid less than males and that the differential be different for non-whites (*non-wh*). In which case we may want to specify a model:

$$\ln(w_i) = \alpha + \beta_1 Female_i + \beta_2 non-wh_i + \beta_3 (Female_i \times non-wh_i) + \varepsilon_i \qquad (8)$$

In which we have 4 alternative cases

1. Male-white

$$\ln(w_i) = \alpha + \beta_1 .0 + \beta_2 .0 + \beta_3 .0.0 + \varepsilon_i$$

$$E[\ln(w_i)] = \alpha \quad (9a)$$

2. Female-white

$$\ln(w_i) = \alpha + \beta_1.1 + \beta_2.0 + \beta_3.1.0 + \varepsilon_i$$

$$E[\ln(w_i)] = \alpha + \beta_1 \quad (9b)$$

3. Male-Non-white

$$\ln(w_i) = \alpha + \beta_1.0 + \beta_2.1 + \beta_3.0.1 + \varepsilon_i$$

$$E[\ln(w_i)] = \alpha + \beta_2 \quad (9c)$$

4. Female-Non-white

$$\ln(w_i) = \alpha + \beta_1.1 + \beta_2.1 + \beta_3.1.1 + \varepsilon_i$$

$$E[\ln(w_i)] = \alpha + \beta_1 + \beta_2 + \beta_3 \quad (9d)$$

Comparing equation (9b) to equation (9a):

$$E\left[\ln(w_i)^{F,Wh}\right] - E\left[\ln(w_i)^{M,Wh}\right] = \beta_1 \quad (10a)$$

that is, β_1 = Proportionate change in expected wages for females compared to males, given they are a white individual.

Comparing equation (9c) to equation (9a):

$$E\left[\ln(w_i)^{M,Non-Wh}\right] - E\left[\ln(w_i)^{M,Wh}\right] = \beta_2 \quad (10b)$$

That is, β_3 = Proportionate change in expected wages for White individuals compared to Non-white individuals given they are males.

Comparing equation (9d) to equation (9c):

$$E\left[\ln(w_i)^{F,Non-Wh}\right] - E\left[\ln(w_i)^{M,Non-Wh}\right] = (\beta_1 + \beta_3) \quad (10c)$$

that is, $\beta_2 + \beta_4$ = Proportionate change in expected wages for females compared to males, given they are a non-white individual

Comparing equation (10c) to equation (10a)

$$\left\{E\left[\ln(w_i)^{F,Non-Wh}\right] - E\left[\ln(w_i)^{M,Non-Wh}\right]\right\} - \left\{E\left[\ln(w_i)^{F,Wh}\right] - E\left[\ln(w_i)^{M,Wh}\right]\right\} = \beta_3 \quad (10d)$$

That is, β_4 = Additional proportionate change in expected wages for females (compared to males) who are a non-white individual compared to a white individual.

Alternatively we could write this as:

	Male	Female	Gender effect	
White	α	$\alpha + \beta_1$	β_2	β_3
Non-white	$\alpha + \beta_2$	$\alpha + \beta_1 + \beta_2 + \beta_3$	$\beta_1 + \beta_3$	
Ethnicity effect	β_2	$\beta_2 + \beta_3$		
	β_3			

This is an example of a difference-in-difference model as the coefficient on the interaction term is interpreted as the difference between two differences (see equation (10d)).

Suppose we know that average salaries listed below:

	Male	Female	Gender effect	
White	15.09	12.58	-2.51	0.09
Non-white	14.92	12.50	-2.42	
Ethnicity effect	-0.17	-0.08		
	0.09			

Then using these numbers and estimating equation (8) by OLS would yield:

$a = 15.09$, $b_1 = -2.51$, $b_2 = -0.17$ and $b_3 = +0.09$.