Week 3. Dummy Variables and Interaction Effects

1. Simple Regression
$$y_i = b_0^* + b_1^* x_{1i} + \varepsilon_i \tag{1}$$
where y = wage and x_1 = years of schooling.

   Let x_2 be a dummy variable that indicates gender. 1 if female and 0 if male.

   From equation 1 in which x_2 is not added, b_1^* refers to the effect of schooling on wage for males and females combined, while b_0^* refers to the intercept (for both groups combined).

2. Regression model with a dichotomy variable (Female)

   x_2 could be added to equation 1 to assess the effect of being female as follows.
$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \varepsilon_i \tag{2}$$
b_0: intercept for males
$b_0 + b_2$: intercept for females
b_1: effect of schooling for both males and females.

   Therefore,

   For male, the prediction becomes $\hat{y}_i = b_0 + b_1 x_{1i}$
For female, it becomes $\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i}$

   This model contrains the effect of schooling on wage to be the same for males and females. However, the model allows for and estimates the effect of being female; b_2 is the wage (dis)advantage for females. Geometrically, b_2 refers to the distance of the female regression line below the male regression line.

3. Significance of the female effect

   To test whether the effect of being female is statistically significant, we would do a t-test where,
$H_0 : b_2 = 0$
$H_A : b_2 \neq 0$

4. Categorical variables

   In general r categories can be handled with $(r - 1)$ dummy variables because the excluded (or reference) category gets the intercept for the overal regression equation.

Suppose that we wish to distinguish between 4 racial groups: whites; blacks; Asians; and others. We would create 3 dummy variables.

$WHT_i = 1$ if white, 0 otherwise
$BLK_i = 1$ if black, 0 otherwise
$ASN_i = 1$ if Asian, 0 otherwise

$$y = b_0 + b_1 x_1 + b_2 WHT_i + b_3 BLK_i + b_4 ASN_i + e \tag{3}$$

b_0: intercept for other races
$b_0 + b_2$: intercept for whites
$b_0 + b_3$: intercept for blacks
$b_0 + b_4$: intercept for Asians

This model specifies four regression lines (one for each group) even though there are three dummy variables because the intercept for the reference group (i.e., excluded category from dummy variables = other races) is b_0.

The t-test of, for example, $H_0 : b_2 = 0$ is a test of whether whites differ from other races. A t-test of any of the (r-1) coefficients is a test of whether that category differs from the excluded or reference category.

5. T-test between dummy variables

To test whether whites differ from blacks we could run the model again but let whites (or blacks) be the reference category as follows:

$$y = b'_0 + b_1 x_1 + b'_3 BLK_i + b'_4 ASN_i + b'_5 OTH_i + e \tag{4}$$

where $OTH_i = 1$ if Other races, otherwise 0.

Or we could do a t-test as follows:

$$t_{b_2 - b_3} = \frac{b_2 - b_3}{\sqrt{s^2_{b_2} + s^2_{b_3}}}$$

where s_{b_2} is the standard error of b_2 and s_{b_3} is the standard error of b_3. Note that this t-test has the same structure of t-test of mean difference between two indepependent samples that we learned in week 4.

Which categry we choose for a reference group does not affect the estimates of relative differences between cateogires. For example, b_2 in equation 3, which is a gap between whites and other races, is equal to $-(b'_5)$ in equation 4. For another example, the gap between whites and Asians in equation 3 ($= b_2 - b_4$) is equal to b'_4 in equation 4. For another example, $b_3 - b_4$ in equaltion 3 is identical to $b'_3 - b'_4$ in equation 4.

6. Interaction between slope and dummy variables

We might also be interested in investigating whether the effect of schooling on wage differs by gender:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + e \tag{5}$$

where $x_{3i} = x_{1i} \times x_{2i}$.

In other words, x_{3i} is an interaction term between x_1 and x_2.

By including x_3 into the model we can test whether or not the female regression line has not only a different intercept but also a different slope as well. A different slope would indicate that the effect of schooling on wage differs by gender.

b_0: intercept for males
b_1: slope for males (i.e., effect of schooling on wage for males)
$b_0 + b_2$: intercept for females
$b_1 + b_3$: slope for females (i.e., effect of schooling on wage for females)

To test whether the slopes differ, we will do a t-test for $H_0 : b_3 = 0$.

Because of the interaction, there is no one effect of gender in equation 5. Rather, there are many different effects of gender depending on the value of x_1. In other words, the effect of being female is dependent on the amount of wage (i.e., x_1).

Effect of being females on wage $= \dfrac{\partial y}{\partial x_2} = b_2 + b_3 x_1$.

Likewise, the effect of schooling is dependent on gender.

Effect of schooling $= \dfrac{\partial y}{\partial x_1} = b_1 + b_3 x_2$

In equation 5, to test whether the effect of being female is statistically zero, we need to do F-test that $H_0 : b_2 = b_3 = 0$. Unless the combined effect of b_2 and b_3 is zero, the effect of being female is significant.