Notes on the reading "Dummy variables and their interactions in regression analysis: examples from research on body mass index".

Major learning outcome from this reading and class: To use dummy variables and their interactions, and to interpret the statistical results adequately.

Data base structure for the example of BMI:

- Three national surveys containing data of 3,323 individuals ages 18-70 in the Netherlands
- Based on self-reported height and weight
- Three years: 2000, 2005 and 2011

Start by studying differences in BMI between males and females, and end our contribution with full-fledged regression models with several explanatory variables and their interactions.

Example #1: BMI Example testing gender differences
(Review of dummy variables with two categories)

Let us start with a simple regression of BMI on gender:

$$PRF: BMI_i = \beta_0 + \beta_1 female + e_i$$

$$SRF: \widehat{BMI}_i = \hat{\beta}_0 + \hat{\beta}_1 female$$

$$\widehat{BMI_i} = 25.23 - 0.51 female$$

Notice (refer to Table 1 in the reading):

- The null hypothesis is that β_1 is zero, as we are testing the alternative hypothesis that β_1 is lower than 0.
- In our sample, the mean BMI difference male-female amounts to -0.51.
- Next, we have to calculate the probability to find -0.51 or even a more negative value in the sample under the null hypothesis (β_1 = 0), i.e. no BMI gender gap in the target population.
- This probability (or p- value) turns out to be lower than .01 (t=-3.82), which is lower than standard test criteria of .05 (corresponding to t=1.96)
- Therefore, we call the outcome of -0.51 statistically significant: it is beyond reasonable doubt that the mean BMI in Dutch adult females is lower than in males.

Example #2: BMI Example testing differences between low, middle and high educated (Review of dummy variables with more than two categories)

To avoid dummy variable trap, we will define “low education” to be the reference category, and include dummy variables for “middle” and “high” education.

$$PRF: BMI_i = \beta_0 + \beta_1 middle + \beta_2 high + e_i$$

$$SRF: \widehat{BMI}_i = \hat{\beta}_0 + \hat{\beta}_1 middle + \hat{\beta}_2 high$$

$$\widehat{BMI_\imath} = 26.12 - 1.18\, middle - 1.83\, high$$

Notice (refer to Table 2 in the reading):

- The low educated (reference category) in our sample have on average an estimated BMI of 26.12.
- The mean BMI of the middle educated is 1.18 points lower (24.94).
- The highly educated have on average the lowest BMI: 1.83 points lower than the mean BMI in the low educated (24.29 and that is).
- These differences (-1.18 and -1.83) both are statistically significant (p-values below .01), as their t-values (-6.76 and -10.19) are both higher than the threshold 1.96

We might also wonder if the difference between high educated and middle educated (which is 24.29 – 24.94 = -.65) is also statistically significant. We can do this by including a parameter for low educated in the model, turning middle education into the reference category:

$$PRF: BMI_i = \beta_0 + \beta_1 low + \beta_2 high + e_i$$

$$SRF: \widehat{BMI}_i = \hat{\beta}_0 + \hat{\beta}_1 low + \hat{\beta}_2 high$$

$$\widehat{BMI_\imath} = 24.94 + 1.18\, low - 0.65\, high$$

Notice (refer to Table 3 in the reading):

- From our first specification in the previous page we knew all the differences between BMI by education attainment:
  - We knew the average BMI of middle educated (reference category in the latter specification): 24.94
  - We knew low educated have a higher BMI than middle, averaging 1.18
  - We knew the difference between middle and high was -0.65
- The specification above only adds the statistics significance of the difference in mean BMI between middle and high educated (-0.65), which we now know is highly significant (p<0.01), as the coefficient has a t=4.41 – well beyond the threshold for 95% significance (t=1.96) and even 99% significance (2.595).

Example #3: Multiple regression model with gender, age, education, number of children.

Here we have a more complex model in which we test whether: a) females on average have a lower BMI than men; b) the middle and higher educated have a lower BMI than the low educated; c) the average BMI has risen since the year 2000; d) the mean BMI rises when someone has children and e) BMI on average rises with age. In our previous example, when we estimated the gender BMI gap (-0.51) we did not account for the fact the most of the time females have a lower educational attainment than men. Here we are aiming to estimating the BMI gender gap, controlling for education, time trend, age, and number of children.

$$\begin{aligned} PRF\colon BMI_i = \beta_0 &+ \beta_1 female + \beta_2 middle + \beta_3 high + \beta_4 onechild + \beta_5 two\ children \\ &+ \beta_6 three\ children + \beta_7 four\ or\ more\ children + \beta_8 2005 + \beta_9 2011 \\ &+ \beta_{10} \ln{(age)} + e_i \end{aligned}$$

$$\begin{aligned} SRF\colon \widehat{BMI_i} = \widehat{\beta_0} &+ \widehat{\beta_1} female + \widehat{\beta_2} middle + \widehat{\beta_3} high + \widehat{\beta_4} onechild + \widehat{\beta_5} two\ children \\ &+ \widehat{\beta_6} three\ children\ + \widehat{\beta_7} four\ or\ more\ children + \widehat{\beta_8} 2005 + \widehat{\beta_9} 2011 \\ &+ \widehat{\beta_{10}} \ln{(age)} \end{aligned}$$

$$\begin{aligned} SRF\colon \widehat{BMI_i} = 23.67 &- 0.58\, female - 0.70\, middle - 1.42\, high + 0.97\, onechild \\ &+ 0.64\, two\ children\ + 0.83\, three\ children + 0.90\, four\ or\ more\ children \\ &+ 0.21\ (2005) + 0.40\ (2011) + 1.95 \ln{(age)} \end{aligned}$$

Notice (refer to Table 4 in the reading):

- After controlling for other relevant factors, females have an estimated mean BMI that is -0.58 points lower than males (as compared to -0.51 in Example #1)
- Respondents with one child have on average a mean BMI that is almost 1 full point higher (0.97) compared to the childless, after taking into account the relevant other factors. Noticeably, this difference does not grow larger with more children.
- Mean BMI has risen over the years in the Netherlands: Compared to the year 2000, mean BMI was 0.21 points higher in 2005 (although the coefficient is not significant as t=1.39 corresponding to p=0.16), and in 2011 it was even .40 points higher.
- The $\widehat{\beta_{10}}$ (coefficient of log-transformed age) is positive and significant, indicating that 1% increase in age, produces a 0.0195 increase in BMI.
- Note that the reading uses age in logs to reflect the possibility that this increase flattens off slightly as one is older.
- This is a rare example of level-log model: According to our rules of thumbs, there is not much reason to use logs here, they could have been substituted for age and age-squared, and the interpretation would have been (hard to calculate but) more natural: With every year of age, BMI increases on average…
- The constant in this specification takes the value 23.67: This is the estimated mean BMI with all variables in the regression model fixed at value 0. In our case this is the estimated mean BMI for males, with a low education, interviewed in 2000,

with no children and at age 18 (because the sample starts at 18 and age was rescaled so that 18=0)

Example #4: Do educational effects differ between males and females?

We want to account for the possibility that the effects of education on BMI are different for woman than for men. We might have some priors on what that might be the case – previous studies, educated guesses, or any conjecture derived from our expertise on the topic – and want to test this hypothesis.

With the toolkit we have revised here, one option you might think of is running the regression of BMI and education for two groups: male and female. Again, to avoid dummy variable trap, we will define "low education" to be the <u>reference category</u>, and have dummy variables for "middle" and "high" education.

$$PRF: BMI_i = \beta_0 + \beta_1 middle + \beta_2 high + e_i$$

$$SRF: \widehat{BMI}_i = \hat{\beta}_0 + \hat{\beta}_1 middle + \hat{\beta}_2 high$$

$$\text{Male: } \widehat{BMI_\iota} = 26.07 - 0.82\, middle - 1.37\, high$$

$$\text{Female: } \widehat{BMI_\iota} = 26.16 - 1.47\, middle - 2.29\, high$$

As can be ascertained from Table 5 in the text, all coefficients of the group regressions are highly significant. Some noteworthy feature of our results up to here:

- At low education level (reference category), there seems to be little differences between the BMI of males and females in our sample (26.07 vs. 26.16)
- At middle education level, the mean value for male is 25.25 (26.07-0.82) is higher than that of females 24.69 (26.16-1.47).
- At high education level, the mean value for male is 24.70 (26.07-1.37) is higher than that of females 23.87 (26.16-1.47).

The only thing we are missing here now is if these differences we observe between the BMI of males and females by education level (0.09 for low, -0.56 for middle, and -0.83 for high) are significant or not. In order to test that, we will use interactions between the female dummy and the education dummy, in the following way:

$$PRF: BMI_i = \beta_0 + \beta_1 female + \beta_2 middle \\ + \beta_3 high + \beta_4 female * middle + \beta_5 female * high + e_i$$

$$SRF: \widehat{BMI}_i = \widehat{\beta_0} + \widehat{\beta_1} female + \widehat{\beta_2} middle + \widehat{\beta_3} high \ + \ \widehat{\beta_4} female * middle \\ + \widehat{\beta_5} female * high$$

$$SRF: \widehat{BMI}_i = 26.07 + 0.09 female - 0.82\, middle - 1.37\, high - \ 0.65 female \\ * middle - 0.92 female * high$$

Notice that this equation for male (female=0) is identical to the just-for-male regression we ran above:

$$SRF: \widehat{BMI}_i = 26.07 - 0.82\, middle - 1.37\, high$$

Likewise, the equation for male (female=1) is identical to the just-for-female regression we ran above:

$$SRF: \widehat{BMI}_i = 26.07 + 0.09 - 0.82\, middle - 1.37\, high - \ 0.65 middle - 0.92 high$$

$$SRF: \widehat{BMI}_i = (26.07 + 0.09) - (0.82 + 0.65) middle - (1.37\, high + 0.92) high$$

$$SRF: \widehat{BMI}_i = 26.16 - 1.47 middle - 2.29 high$$

Notice (refer to Table 6 in the reading):

- The BMI gender gap for the middle educated (-0.56) is the sum of the BMI gender gap among the low educated (β_1 =0.09) and the interaction female*middle ($\beta_4 = -0.65$).
- This value indicates how much the BMI gender gap in general (0.09) differs from the gap among the middle educated (-0.65), which is exactly what we wanted to know.
- Notice that the differences in mean BMI between groups (male and female) are not statistically different from zero, *once we control for the difference in education by gender*.
- Notice that the dummy for the interaction term *female*middle* is not significant at the 95% level (t=1.86 falls short of 1.96), indicating that BMI differences across genders with middle education is not statistically different from zero at 95% level (it is significant at the 10% level, or t=1.645)
- In a similar way, the BMI gender gap for the highly educated (-0.83) is the sum of the BMI gender gap among the low educated (β_1 =0.09) and the interaction female*high ($\beta_4 = -0.92$).
- Given the large t-statistic of the interaction *female*high* (t=2.57), we can reject the null hypothesis of no differences in the means of males and females among the high educated.