# HANDOUT 9 – INTERACTIONS

## AGENDA

Non-linear regression functions

- Introduction
- Dummy variable review: The case of BMI
- Dummy-dummy interactions: The case of BMI
- Continuous-dummy interactions: The case of immigrants' wage gap
- Takeaways
- Vocabulary

## BIBLIOGRAPHY FOR TODAY´S CLASS

- Te Grotenhuis, M., Thijs, P. (2015). Dummy variables and their interactions in regression analysis: examples from research on body mass index. (*)
- Stock and Watson (2007), 8.3 (**)

## INTRODUCTION

We continue our study of nonlinear relationships:

- Last time, we allowed the predicted change in Y associated with a change in $X_1$ to depend upon the value of $X_1$
- Today we introduce nonlinearity in another way: We allow the predicted change in Y associated with a change in $X_1$ to depend on another variable $X_2$

## DUMMY VARIABLE REVIEW: THE CASE OF THE BMI

We are first trying to understand the associated between gender and BMI:

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

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

**Table 1** The estimated BMI means for males (*a*) and the BMI gender gap (*b*)

| | coefficients | standard error | t-value | p-value (2-tailed) |
|---|---|---|---|---|
| a | 25.23 | .10 | 260.90 | <.01 |
| b *female* (0=male, 1=female) | -.51 | .13 | -3.82 | <.01 |

We then hypothesize that education might have something to do with BMI:

$$PRF: BMI_i = \beta_0 + \beta_1 middle + \beta_2 high + e_i$$
$$SRF: \widehat{BMI}_i = \hat{\beta}_0 + \hat{\beta}_1 middle + \hat{\beta}_2 high$$

**Table 2** The estimated BMI means for low education (a) and the mean differences with middle ($b_1$) and high education ($b_2$)

| | coefficients | standard error | t-value | p-value (2-tailed) |
|---|---|---|---|---|
| a (low) | 26.12 | .14 | 183.17 | <.01 |
| $b_1$ (middle) | -1.18 | .17 | -6.76 | <.01 |
| $b_2$ (high) | -1.83 | .18 | -10.19 | <.01 |

DUMMY-DUMMY INTERACTIONS: THE CASE OF BODY MASS INDEX (BMI)

We want to test if the effects of education on BMI are different by gender

- Is the association between BMI and gender the same for different levels of education attainment?

One option we have is to run two different regressions for male and female:

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

| | coefficients | standard error | t-value | p-value (2-tailed) |
|---|---|---|---|---|
| **MALES** | | | | |
| *a (constant)* | 26.07 | .18 | 145.38 | .00 |
| low | reference | | | |
| middle | -.82 | .22 | -3.65 | .00 |
| high | -1.37 | .23 | -6.09 | .00 |
| **FEMALES** | | | | |
| *a (constant)* | 26.16 | .22 | 120.25 | .00 |
| low | reference | | | |
| middle | -1.47 | .26 | -5.59 | .00 |
| high | -2.29 | .28 | -8.32 | .00 |

$$\text{Male: } \widehat{BMI}_i = 26.07 - 0.82\ middle - 1.37\ high$$

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

The only thing we are missing on the table above is if the 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.

To test if the differences 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 we can use an interaction

We will define a variable called an interaction:

- The multiplicative product of two explanatory variables.
- Can be added to our usual regressions.

In our particular example, to find out if the differential impact of gender on BMI changes with educational attainment, we need to define an interaction between the female dummy with the education dummies:

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

| | coefficients | standard error | t-value | p-value (2-tailed) |
|---|---|---|---|---|
| **Main effects** | | | | |
| *a (constant)* | 26.07 | .20 | 128.18 | .00 |
| female ($b_1$) | .09 | .28 | .31 | .75 |
| low | reference | | | |
| middle ($b_2$) | -.82 | .25 | -3.22 | .00 |
| high ($b_3$) | -1.37 | .26 | -5.37 | .00 |
| **Interaction effects** | | | | |
| female * low | reference | | | |
| female * middle ($b_4$) | -.65 | .35 | -1.86 | .06 |
| female * high ($b_5$) | -.92 | .36 | -2.57 | .01 |

How can we reconcile the coefficients $\beta_1$ and $\beta_4$ with the difference in BMI between middle educated man and female (-0.56) that we got in our separate regressions?

Is there a significant difference between average BMI of male and females at all different levels of educational attainment?

How can we interpret $\beta_4$?

How can we interpret $\beta_5$?

What happens to our specification with interactions…

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

If the subject is male?

If the subject is female?

What is the predicted BMI for each type of individual?

| | LOW EDUCATION | MIDDLE EDUCATION | HIGH EDUCATION | DIFFERENCE MIDDLE-LOW | DIFFERENCE HIGH-LOW |
|---|---|---|---|---|---|
| **MALE** | | | | | |
| **FEMALE** | | | | | |
| **DIFFERENCE (F-M)** | | | | | |

CONTINUOUS-DUMMY INTERACTIONS: THE CASE OF IMMIGRANTS' WAGE GAP

Consider this regression of wages on years of education (beyond 8th grade) and an indicator for whether an individual is an immigrant:

$$wage = \beta_0 + \beta_1 immigrant + \beta_2 educ + \varepsilon$$

Using OLS, we can estimate this with data from the 2016 American Community Survey on 30-50 years-old MA residents (N=17,288)

```
. reg incwage educ immigrant, robust noheader
```

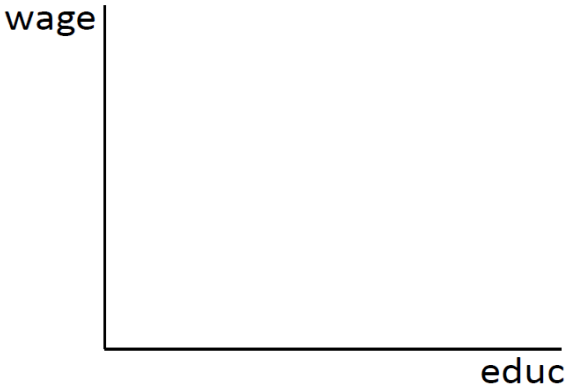
| incwage | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| educ | 7816.583 | 184.9456 | 42.26 | 0.000 | 7454.071 | 8179.095 |
| immigrant | -3863.488 | 1168.296 | -3.31 | 0.001 | -6153.467 | -1573.509 |
| _cons | 8493.176 | 1142.596 | 7.43 | 0.000 | 6253.573 | 10732.78 |

What's the estimated change in wages associated with one-more year of schooling

- For immigrants?

- For non-immigrants?

How can we represent this graphically?




The assumption that the immigration wage gap is the same for all levels of education may not be realistic.

We want the regression model to allow the **slope** (not just the **constant**) to differ by immigration status.

We can do so by creating an **interaction** between the dummy variable *immigrant* and the continuous variable *educ* (by multiplying the two):

```
. gen immig_educ = immigrant*educ

. list immigrant educ immig_educ in 1/10
```

| | immigrant | educ | immig_educ |
|---|---|---|---|
| 1. | 1 | 8 | 8 |
| 2. | 0 | 8 | 0 |
| 3. | 0 | 10 | 0 |
| 4. | 1 | 8 | 8 |
| 5. | 0 | 6 | 0 |
| 6. | 0 | 8 | 0 |
| 7. | 0 | 8 | 0 |
| 8. | 0 | 10 | 0 |
| 9. | 1 | 10 | 10 |
| 10. | 1 | 8 | 8 |

Notice that I cannot produce *immigrant*educ* by subtracting immigrant from *educ*, or by multiplying neither *educ* or *immigrant* by a constant: *immigrant*educ* is a new variable – this is not a case of multicollinearity.

To see why that interaction variable helps, consider this regression model:

$$wage = \beta_0 + \beta_1 immigrant + \beta_2 educ + \beta_3 immig\_educ + \varepsilon$$

What is the predicted wage change associated with one more year of education for:

- Non-immigrants?

- Immigrants?

What is the interpretation of:

- $\widehat{\beta_1}$:

- $\widehat{\beta_3}$:

Here are the actual results from the data:

```
. reg incwage educ immigrant immig_educ, robust noheader
```

| incwage | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| educ | 8843.983 | 245.2834 | 36.06 | 0.000 | 8363.202 | 9324.763 |
| immigrant | 12218.93 | 1927.633 | 6.34 | 0.000 | 8440.576 | 15997.29 |
| immig_educ | -2575.869 | 365.3809 | -7.05 | 0.000 | -3292.053 | -1859.686 |
| _cons | 1535.086 | 1450.756 | 1.06 | 0.290 | -1308.543 | 4378.715 |

What is the sample regression function for:

- Non-immigrants?

- Immigrants?

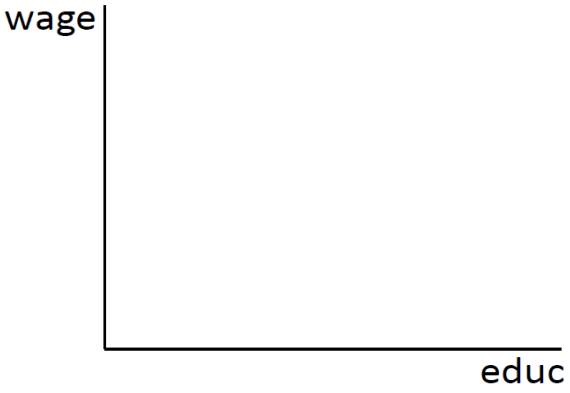
What's the predicted wage change associated one more year of education for

- Non-immigrants?

- Immigrants?

- Which slope is steeper?

How can we represent these findings graphically?




TAKEAWAYS

VOCABULARY

- Interactions: when interaction terms are included as regressors, they allow the regression slope of one variable to depend on the value of another variable.
    - Continuous-dummy: through the use of the interaction term, the regression line relating the dependent variable and the continuous variable can have a slope that depends on the binary variable in three different ways:
        - Different intercept, same slope.
        - Same intercept, different slope.
        - Different intercept and slope.
    - Dummy-dummy: the binary variable interaction regression allows the effect of changing one of the binary independent variables to depend on the value of another binary variable.