API-202B Empirical Methods II

Session #9: Interactions

miguel_santos@hks.harvard.edu
@miguelsantos12

Today´s class: Interactions

- Introduction
- Dummy variable review: The case of BMI
- Dummy-dummy interactions: The case of the BMI
- Continuous-dummy interactions: The case of immigration wage gap
- Takeaways

- 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

The case of Body Mass Index (BMI): Dummy variable review

- 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

The case of Body Mass Index (BMI): Dummy variable review

- 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
- One option 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

Male: $\widehat{BMI_i} = 26.07 - 0.82\, middle - 1.37\, high$

Female: $\widehat{BMI_i} = 26.16 - 1.47\, middle - 2.29\, high$

Dummy-Dummy Interactions: The case of Body Mass Index (BMI)

- 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:

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

Dummy-Dummy Interactions: The case of Body Mass Index (BMI)

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

	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

Dummy-Dummy Interactions: The case of Body Mass Index (BMI)

- How can we reconcile the coefficients β_1 and β_4 with the difference in BMI between middle educated man and female (-0.56) that we got in our separate regressions?
  - 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 (β_4 =-0.65).
- Is there a significant difference between average BMI of male and females at all different levels of educational attainment?
  - Differences in mean BMI between groups of low education (male and female) are not statistically different from zero.
  - 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 gender with middle education is not statistically different from zero at 95% level.
  - Dummy for the interaction term female*high is significant at the 95% level (t=2.57 is higher than 1.96), indicating that BMI differences across genders with high education are statistically significant from zero at 95% level (lower for female)
- How can we interpret β_4?
- Is the differential returns of female with respect to man of same education (middle).
- How can we interpret β_5?
- Is the differential returns of female with respect to man of same education (high).

Dummy-Dummy Interactions: The case of Body Mass Index (BMI)

- 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? (Female=0)
  - It becomes identical to the single regression we run for male (slide 5)

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

- If the subject is female? (Female=1)
  - It becomes identical to the single regression we run for male (slide 5)

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

- The difference is that with the specification with interactions we can test if the differences between male and female are significant at all education levels

Dummy-Dummy Interactions: The case of Body Mass Index (BMI)

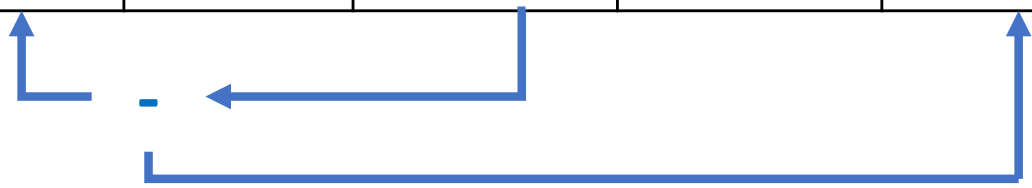
- What is the predicted BMI for each type of individual?

	LOW EDUCATION	MIDDLE EDUCATION	HIGH EDUCATION	DIFFERENCE MIDDLE-LOW	DIFFERENCE HIGH-LOW
MALE	26.07	25.25	24.70	-0.82	-1.37
FEMALE	26.16	24.69	23.87	-1.47	-2.29
DIFFERENCE (F-M)	0.09	-0.56	-0.83	-0.65	-0.92

Dummy-Dummy Interactions: The case of Body Mass Index (BMI)

- What is the predicted BMI for each type of individual?

	LOW EDUCATION	MIDDLE EDUCATION	HIGH EDUCATION	DIFFERENCE MIDDLE-LOW	DIFFERENCE HIGH-LOW
MALE	26.07	25.25	24.70	-0.82	-1.37
FEMALE	26.16	24.69	23.87	-1.47	-2.29
DIFFERENCE (F-M)	0.09	-0.56	-0.83	-0.65	-0.92



Continuous-Dummy Interactions: Wages, migrants and education

- 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 year old MA residents (N=17,288)

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

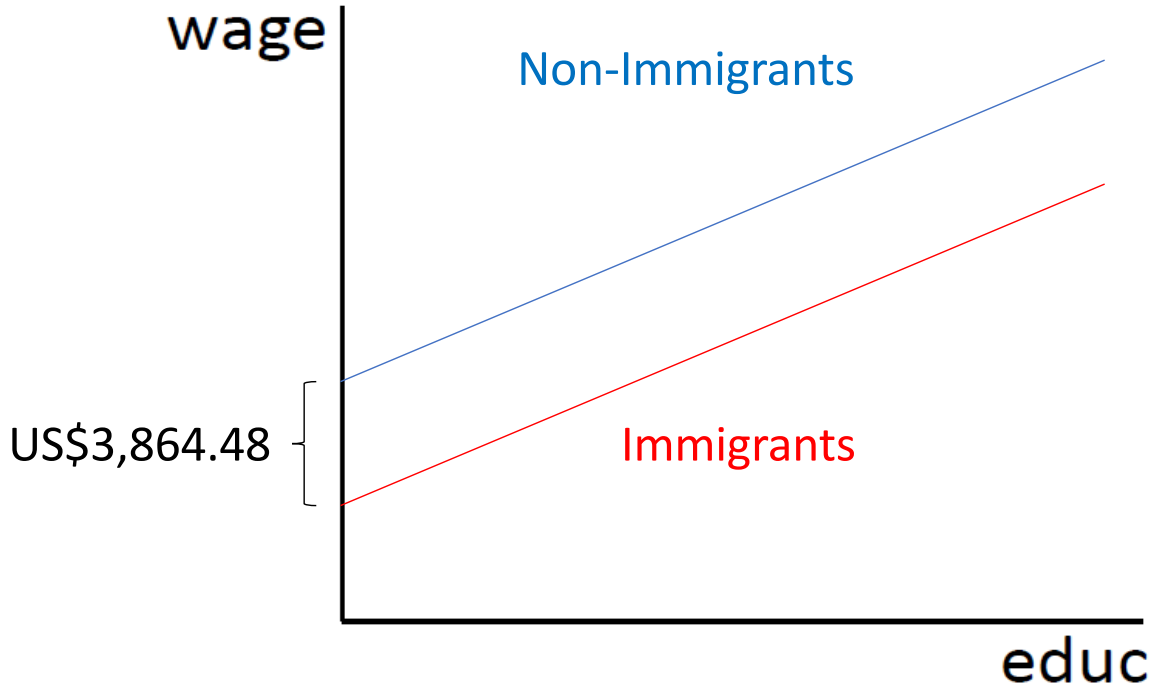
incwage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	7816.583	184.9456	42.26	0.000	7454.071	8179.095
immigrant	-3863.488	1168.296	-3.31	0.001	-6153.467	-1573.509
_cons	8493.176	1142.596	7.43	0.000	6253.573	10732.78

Continuous-Dummy Interactions: Wages, migrants and education

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

It´s the same, regardless of the immigrant status. The specification without interactions give is the average mean value of wage associated with every additional year of education.

- How would a graphic representation of our regression model look?



Continuous-Dummy Interactions: Wages, migrants and education

- 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

Continuous-Dummy Interactions: Wages, migrants and education

- 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? (immigrant=0)

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

  - Immigrants? (immigrant=1)

$$wage = \beta_0 + \beta_1(1) + \beta_2 educ + \beta_3(1) educ + \varepsilon$$

$$wage = \underbrace{(\beta_0 + \beta_1)}_{\text{Different constant or intercept}} + \underbrace{(\beta_2 + \beta_3)}_{\text{Different slope}} educ + \varepsilon$$

- What is the interpretation of:
  - $\widehat{\beta_1}$: Incremental base (associated to zero education) for immigrants
  - $\widehat{\beta_3}$: Differential slope for every additional year of education for immigrants

Continuous-Dummy Interactions: Wages, migrants and education

- 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

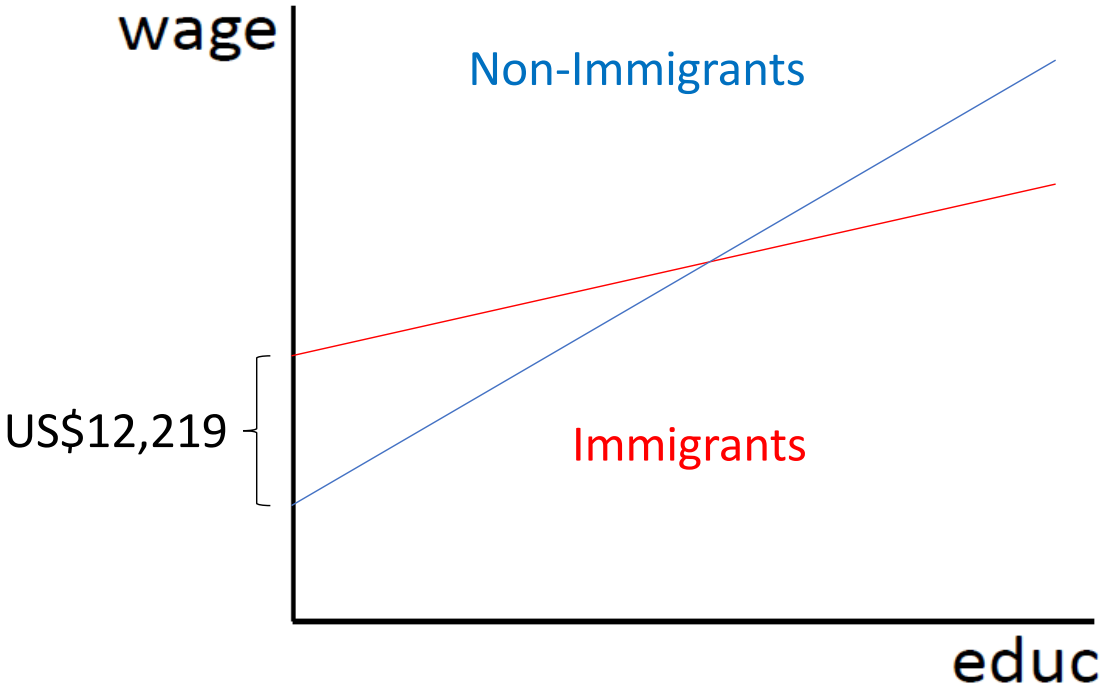
- What is the sample regression function for
  - Non-immigrants? $wage = 1{,}535 + 8{,}844\, educ + \varepsilon$
  - Immigrants? $wage = (1{,}535 + 12{,}219) + (8{,}844 - 2{,}576)\, educ + \varepsilon$

$$wage = 13{,}754 + 6{,}269\, educ + \varepsilon$$

Continuous-Dummy Interactions: Wages, migrants and education

- What's the predicted wage change associated one more year of education for
  - Non-immigrants? 8,844
  - Immigrants? $8{,}844 - 2{,}576 = 6{,}269$

- Which slope is steeper?
  - Non immigrants!
  - 8,844 > 6,269
- How can we represent these findings graphically?



Takeaways

- Interactions allows us to explore if the predicted change in Y associated with a change in X_1 depends on another variable X_2 (change meaning slopes)
- The statistical significance of the interaction coefficient reveals whether the relationship of interest varies between the groups being distinguished
- Interactions are often used to explore **heterogeneity**, i.e. whether relationships between two variables differ by gender, income, race, etc.
- In case you forget how to interpret interaction coefficients, try predicting outcomes for various groups in your data one at a time. This may help clarify for you what each coefficient represents
- We will also see interactions used to implement a method of causal inference called **differences-in-differences**