Eco 213: Basic Data Analysis and Econometrics
Lecture 4 : Dummy Variables
February 23, 2019

Dr. Garima Malik
Department of Economics
Shiv Nadar University

Dummy Variables

- Question: How can we incorporate nominal variables (e.g., race, gender) into regression?
- Option 1: Analyze each sub-group separately
  - Generates different slope, constant for each group
- Option 2: Dummy variables
  - "Dummy" = a dichotomous variables coded to indicate the presence or absence of something
  - Absence coded as zero, presence coded as 1.

Dummy Variables

- Strategy: Create a separate dummy variable for **all** nominal categories
- Ex: Gender – make female & male variables
  - DFEMALE: coded as 1 for all women, zero for men
  - DMALE: coded as 1 for all men
- Next: Include **all but one** dummy variables into a multiple regression model
  - If two dummies, include 1; If 5 dummies, include 4.

Dummy Variables

- Question: Why can't you include DFEMALE and DMALE in the same regression model?
- Answer: They are perfectly correlated (negatively): r = -1
  - Result: Regression model "blows up"
- For any set of nominal categories, a full set of dummies contains redundant information
  - DMALE and DFEMALE contain same information
  - Dropping one removes redundant information.

Dummy Variables: Interpretation

- Consider the following regression equation:

$$Y_i = a + b_1 INCOME_i + b_2 DFEMALE_i + e_i$$

- Question: What if the case is a male?
- Answer: DFEMALE is 0, so the entire term becomes zero.
  - Result: Males are modeled using the familiar regression model: $a + b_1X + e$.

Dummy Variables: Interpretation

- Consider the following regression equation:

$$Y_i = a + b_1 INCOME_i + b_2 DFEMALE_i + e_i$$

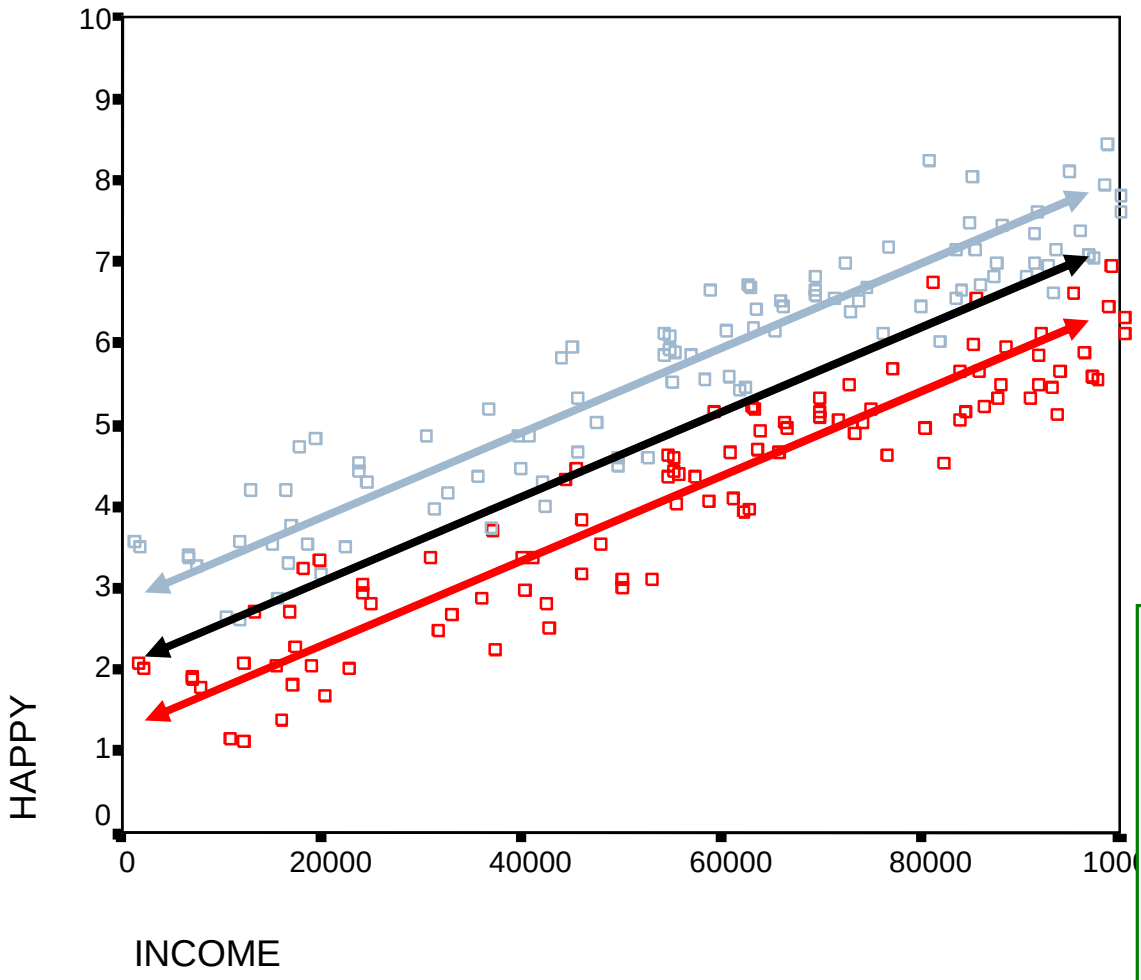
- Question: What if the case is a female?
- Answer: DFEMALE is 1, so $b_2(1)$ stays in the equation (and is added to the constant)
  - Result: Females are modeled using a different regression line: $(a+b_2) + b_1X + e$
  - Thus, the coefficient of b_2 reflects difference in the **constant** for women.

Dummy Variables: Interpretation

- Remember, a different constant generates a different line, either higher or lower
  - Variable: DFEMALE (women = 1, men = 0)
  - A positive coefficient (b) indicates that women are consistently higher compared to men
  - A negative coefficient indicated women are lower
- Example: If DFEMALE coeff = 1.2:
  - “Women are on average 1.2 points higher than men”.

Dummy Variables: Interpretation

- Visually: Women = blue, Men = red



Overall slope for all data points

Note: Line for men, women have same slope… but one is high other is lower. The constant differs!

If women=1, men=0: The constant (a) reflects men only. Dummy coefficient (b) reflects increase for women (relative to men)

Dummy Variables

- What if you want to compare more than 2 groups?
- Example: Race
  - Coded 1=white, 2=black, 3=other
  - Make 3 dummy variables:
  - "DWHITE" is 1 for whites, 0 for everyone else
  - "DBLACK" is 1 for Af.Am., 0 for everyone else
  - "DOTHER" is 1 for "others", 0 for everyone else
- Then, include **two** of the three variables in the multiple regression model.

Dummy Variables: Interpretation

- Ex: Job Prestige

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	9.666	1.672		5.780	.000
	EDUC	2.476	.111	.517	22.271	.000
	INCOM16	6.282E-02	.397	.004	.158	.874
	DBLACK	-2.666	1.117	-.055	-2.388	.017
	DOTHER	1.114	1.777	.014	.627	.531

a. Dependent Variable: PRESTIGE

- Negative coefficient for DBLACK indicates a lower level of job prestige compared to whites
  - T- and P-values indicate if difference is significant.

Dummy Variables: Interpretation

- Comments:
- 1. Dummy coefficients shouldn't be called slopes
  - Referring to the "slope" of gender doesn't make sense
  - Rather, it is the difference in the **constant** (or "level")
- 2. The contrast is **always** with the nominal category that was **left out** of the equation
  - If DFEMALE is included, the contrast is with males
  - If DBLACK, DOTHER are included, coefficients reflect difference in constant compared to whites.

Interaction Terms

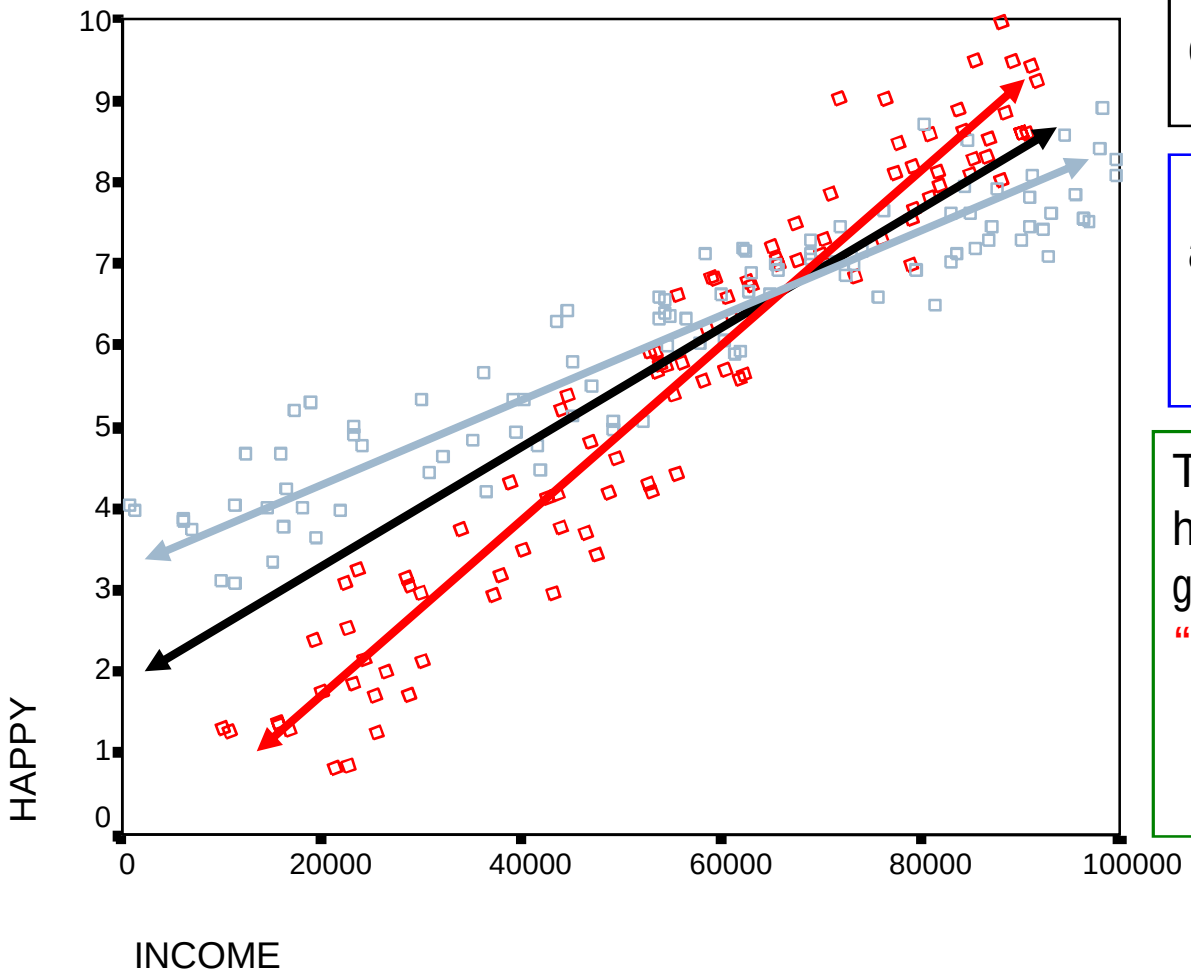
- Question: What if you suspect that a variable has a totally different slope for two different sub-groups in your data?
- Example: Income and Happiness
  - Perhaps men are more materialistic -- an extra dollar increases their happiness a lot
  - If women are less materialistic, each dollar has a smaller effect on income (compared to men)
- Issue isn't men = "more" or "less" than women
  - Rather, the slope of a variable (income) differs across groups

Interaction Terms

- Issue isn't men = "more" or "less" than women
  - Rather, the slope of a variable coefficient (for income) differs across groups
- Again, we want to specify a different regression line for each group
  - We want lines with different slopes, not parallel lines that are higher or lower.

Interaction Terms

- Visually: Women = blue, Men = red



Overall slope for all data points

Note: Here, the slope for men and women differs.

The effect of income on happiness (X1 on Y) varies with gender (X2). This is called an "interaction effect"

Interaction Terms

- Interaction effects: Differences in the relationship (slope) between two variables for each category of a **third variable**
- Option #1: Analyze each group separately
- Option #2: Multiply the two variables of interest: (DFEMALE, INCOME) to create a new variable
  - Called: DFEMALE*INCOME
  - Add that variable to the multiple regression model.

Interaction Terms

- Consider the following regression equation:

$$Y_i = a + b_1 INCOME_i + b_2 DFEM * INC_i + e_i$$

- Question: What if the case is male?
- Answer: DFEMALE is 0, so b_2(DFEM*INC) drops out of the equation
  - Result: Males are modeled using the ordinary regression equation: $a + b_1X + e$.

Interaction Terms

- Consider the following regression equation:

$$Y_i = a + b_1 INCOME_i + b_2 DFEM * INC_i + e_i$$

- Question: What if the case is male?
- Answer: DFEMALE is 1, so b_2(DFEM*INC) becomes b_2*INCOME, which is added to b_1
  - Result: Females are modeled using a different regression line: a + (b_1+b_2) X + e
  - Thus, the coefficient of b_2 reflects difference in the **slope** of INCOME for women.

Interaction Terms

- Interpreting interaction terms:
- A positive b for DFEMALE*INCOME indicates the slope for income is higher for women vs. men
  - A negative effect indicates the slope is lower
  - Size of coefficient indicates actual difference in slope
- Example: DFEMALE*INCOME. Observed b's:
  - Income: b = .5
  - DFEMALE * INCOME: b = -.2
- Interpretation: Slope is .5 for men, .3 for women.

Interaction Terms

- Continuous variable can also interact
- Example: Effect of education and income on happiness
  - Perhaps highly educated people are less materialistic
  - As education increases, the slope between between income and happiness would decrease
- Simply multiply Education and Income to create the interaction term “EDUCATION*INCOME”
  - And add it to the model

Interaction Terms

- How do you interpret continuous variable interactions?
  - Example: EDUCATION*INCOME: Coefficient = 2.0
- Answer: For each unit change in education, the slope of income vs. happiness increases by 2
  - Note: coefficient is symmetrical: For each unit change in income, education slope increases by 2
  - Dummy interactions result in slopes for each group
  - Continuous interactions result in many slopes
    - Each category of education*income has a different slope.

Interaction Terms

- 1. If you make an interaction you should also include the component variables in the model:
  - A model with "DFEMALE * INCOME" should also include DFEMALE and INCOME
  - There is some debate on this issue… but that is the safest course of action
- 2. Sometimes interaction terms are highly correlated with its components