ECON 340
Economic Research Methods

Div Bhagia

Lecture 19
Categorical Variables & Interaction Terms

Fitting a Line

Linear relationship (with some error):

$$Y = \beta_0 + \beta_1 X + u$$

Taking the conditional expectation:

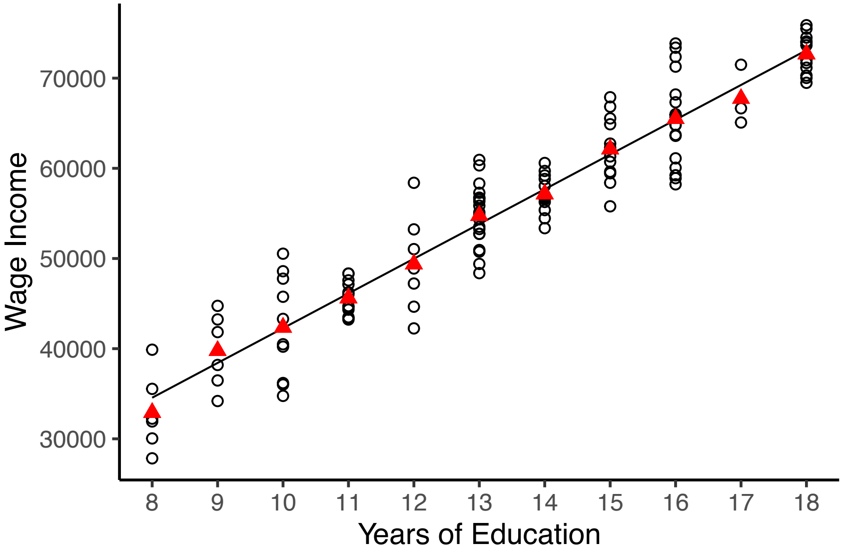
$$E(Y|X) = \beta_0 + \beta_1 X + E(u|X)$$

With $E(u|X) = 0$,

$$E(Y|X) = \beta_0 + \beta_1 X$$

OLS fits a linear line between average Y at each X and X.

Hypothetical Data: $E(wages|educ)$ and $educ$



Dummy Variables

What if the independent variable is a binary variable that takes two values 1 and 0?

$$Y = \beta_0 + \beta_1 D + u$$

Taking conditional expectation (assuming exogeneity):

$$E[Y|D = 1] = \beta_0 + \beta_1 \cdot 1 = \beta_0 + \beta_1$$
$$E[Y|D = 0] = \beta_0 + \beta_1 \cdot 0 = \beta_0$$

So,

$$\beta_1 = E[Y|D = 1] - E[Y|D = 0]$$

ACS Data: Gender Wage Gap

	Wages
Intercept	67,220.17***
	(439.87)
Female	−14,661.12***
	(637.27)
Observations	17,578
R^2	0.03
Note:	*p<0.1; **p<0.05; ***p<0.01

Dummy Variables: Interpretation

As before, to interpret β_1 as the causal impact of gender on wages, we need:

$$E(u|female) = 0$$

Meaning that omitted factors that impact wages are uncorrelated with gender, which implies:

$$\beta_1 = E[wages|female = 1] - E[wages|female = 0]$$

However, even if exogeneity doesn't hold, $\hat{\beta}_1$ still captures the difference in average wages of men and women in our sample.

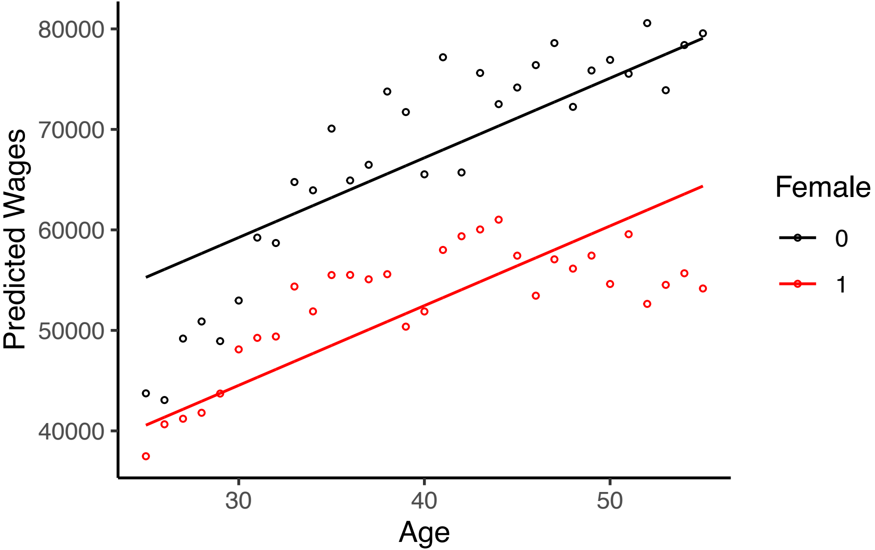
Dummy Variables in Multiple Regression

$$Wages = \beta_0 + \beta_1 Age + \beta_2 Female + u$$

Taking conditional expectation (assuming exogeneity):

$$E[Wages|Age, Female = 1] = (\beta_0 + \beta_2) + \beta_1 Age$$
$$E[Wages|Age, Female = 0] = \beta_0 + \beta_1 Age$$

ACS Data: Wages and Age



Interaction Terms

We can also include interaction terms in our model as follows:

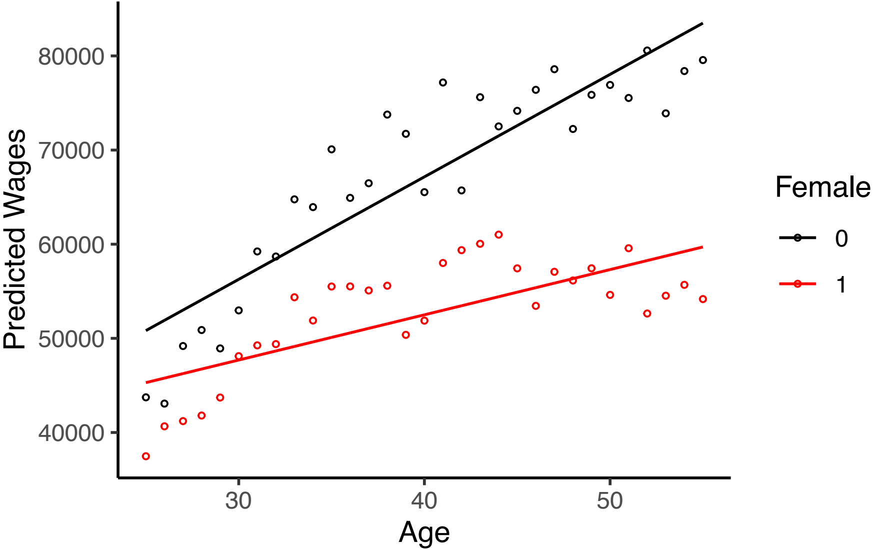
$$Wages = \beta_0 + \beta_1 Age + \beta_2 Female + \beta_3 Female \times Age + u$$

Taking conditional expectation (assuming exogeneity):

$$E[Wages|Age, Female = 1] = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)Age$$
$$E[Wages|Age, Female = 0] = \beta_0 + \beta_1 Age$$

Now the impact of X on Y varies with D.

ACS Data: Wages and Age



Interaction of Two Dummy Variables

$$wages = \beta_0 + \beta_1 Female + \beta_2 Hispanic + \beta_3 Female \times Hispanic + u$$

Average wages for Non-Hispanic Males:

$$E(wages | Hispanic = 0, Female = 0) = \beta_0$$

Average wages for Non-Hispanic Females:

$$E(wages | Hispanic = 0, Female = 1) = \beta_0 + \beta_1$$

Interaction of Two Dummy Variables

$$wages = \beta_0 + \beta_1 Female + \beta_2 Hispanic + \beta_3 Female \times Hispanic + u$$

Average wages for Hispanic Males:

$$E(wages | Hispanic = 1, Female = 0) = \beta_0 + \beta_2$$

Average wages for Hispanic Females:

$$E(wages | Hispanic = 1, Female = 1) = \beta_0 + {\color{red}\beta_1} + \beta_2 + {\color{red}\beta_3}$$

ACS Data: Gender and Ethnicity

	Wages
Intercept	70,179.09***
	(473.52)
Female	−16,046.81***
	(683.42)
Hispanic	−19,367.71***
	(1,211.46)
Female X Hispanic	8,163.75***
	(1,788.04)
Observations	17,578

Variable with Multiple Categories

Five education categories:

$\{$Less than HS, HS Grad, Some College, College Degree, >College$\}$

Add four dummy variables to the regression (why not five?):

$$wages = \beta_0 + \beta_1 HS + \beta_2 SomeCol + \beta_3 Col + \beta_4 MoreThanCol + u$$

Reference category: Less than HS

Coefficients capture the difference between average wages for that category and average wages for *less than HS*.

Variable with Multiple Categories

Education	Wages
Less than HS	36090.83
High School	44546.88
Some College	50182.94
College Degree	71527.75
More than College	87775.73

Variable with Multiple Categories

	Wages
Intercept	36,090.83***
	(1,386.07)
High School	8,456.05***
	(1,496.75)
Some College	14,092.11***
	(1,515.36)
College Degree	35,436.92***
	(1,499.47)
More than College	51,684.90***
	(1,559.17)
Observations	17,578

Binary Dependent Variable

What if we have a binary variable on the left-hand side?

$$emp = \beta_0 + \beta_1 educ + u$$

$$E[emp|educ] = \beta_0 + \beta_1 educ$$

Note that,

$$E[emp|educ] = P(emp = 1|educ) = \beta_0 + \beta_1 educ$$

So, β_1 can be interpreted as the change in the probability of being employed. This is called the Linear Probability Model.