Multiple Linear Regression (Dummy Variable Treatment)

CIVL 7012/8012

In Today's Class

- Recap
- Single dummy variable
- Multiple dummy variables
- Ordinal dummy variables
- Dummy-dummy interaction
- Dummy-continuous/discrete interaction
- Binary dependent variables

Introducing Dummy Independent Variable

- **Qualitative Information**
  - Examples: gender, race, industry, region, rating grade, …
  - A way to incorporate qualitative information is to use dummy variables
  - They may appear as the dependent or as independent variables
- **A single dummy independent variable**

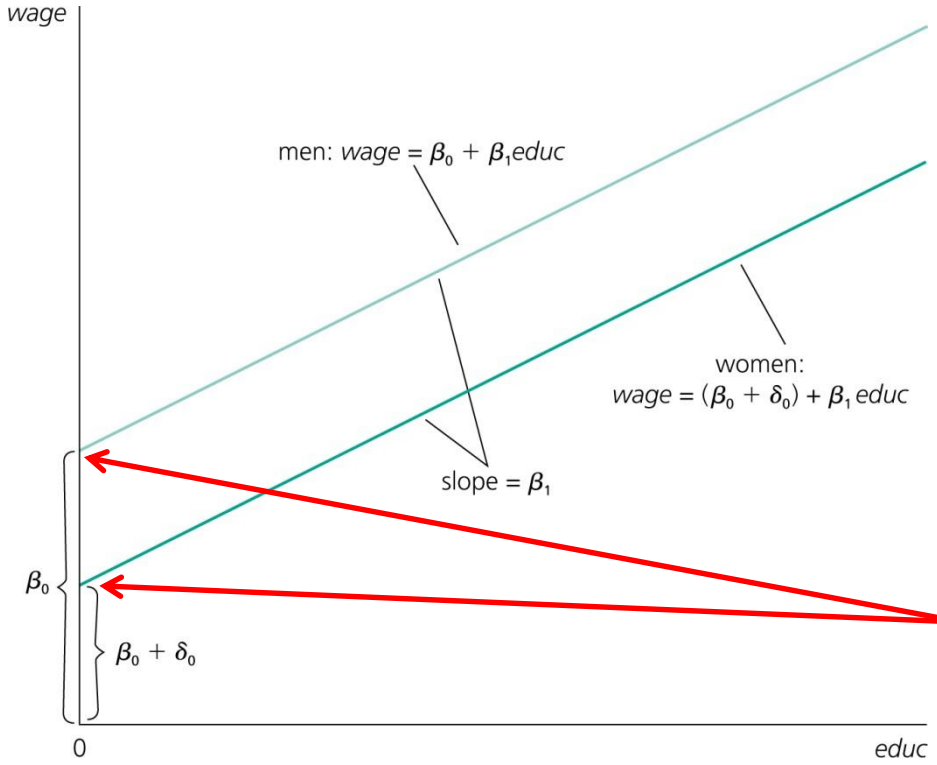
$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u$$

= the wage gain/loss if the person is a woman rather than a man (holding other things fixed)

Dummy variable:
=1 if the person is a woman
=0 if the person is man

Illustrative Example

- **Graphical Illustration**



Alternative interpretation of coefficient:

$$\delta_0 = E(wage|female = 1, educ) - E(wage|female = 0, educ)$$

i.e. the difference in mean wage between men and women with the same level of education.

Intercept shift

Specification of Dummy Variables

- **Dummy variable trap**

This model cannot be estimated (perfect collinearity)

$$wage = \beta_0 + \gamma_0 male + \delta_0 female + \beta_1 educ + u$$

When using dummy variables, one category always has to be omitted:

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u$$ ← The base category are men

$$wage = \beta_0 + \gamma_0 male + \beta_1 educ + u$$ ← The base category are women

Alternatively, one could omit the intercept:

$$wage = \gamma_0 male + \delta_0 female + \beta_1 educ + u$$

Disadvantages:
1) More difficult to test for differences between the parameters
2) R-squared formula only valid if regression contains intercept

Interpretation of Dummy Variables

- **Estimated wage equation with intercept shift**

$$\widehat{wage} = -\underset{(.72)}{1.57} - \underset{(.26)}{1.81}\, female + \underset{(.049)}{.572}\, educ$$

$$+ \underset{(.012)}{.025}\, exper + \underset{(.021)}{.141}\, tenure$$

$$n = 526, R^2 = .364$$

Holding education, experience, and tenure fixed, women earn 1.81$ less per hour than men

(Standard errors in parenthesis)

- **Does that mean that women are discriminated against?**
  - Not necessarily. Being female may be correlated with other productivity characteristics that have not been controlled for.

Model with only dummy variables- (Example-1)

- **Comparing means of subpopulations described by dummies**

$$\widehat{wage} = \underset{(.21)}{7.10} - \underset{(.26)}{2.51}\, female$$

$$n = 526, R^2 = .116$$

Not holding other factors constant, women earn 2.51$ per hour less than men, i.e. the difference between the mean wage of men and that of women is 2.51$.

- **Discussion**
  - It can easily be tested whether difference in means is significant
  - The wage difference between men and women is larger if no other things are controlled for; i.e. part of the difference is due to differences in education, experience and tenure between men and women

Model with only dummy variables- (Example-2)

- **Further example: Effects of training grants on hours of training**

Hours training per employee

Dummy indicating whether firm received training grant

$$\widehat{hrsemp} = \underset{(43.41)}{46.67} + \underset{(5.59)}{26.25}\,grant - \underset{(3.54)}{0.98}\,\log(sales)$$

$$- \underset{(3.88)}{6.07}\,\log(employ),\ n = 105, R^2 = .237$$

- **This is an example of <u>program evaluation</u>**
  - Treatment group (= grant receivers) vs. control group (= no grant)
  - Is the effect of treatment on the outcome of interest <u>causal</u>?

Dependent log(y) and Dummy Independent

- **Using dummy explanatory variables in equations for log(y)**

$$\widehat{\log(price)} = \underset{(.65)}{-1.35} + \underset{(.038)}{.168}\log(lotsize) + \underset{(.093)}{.707}\log(sqrft)$$

$$+ \underset{(.029)}{.027}\,bdrms + \underset{(.045)}{.054}\,colonial$$

Dummy indicating whether house is of colonial style

$$n = 88, R^2 = .649$$

$$\Rightarrow \quad \frac{\partial \log(price)}{\partial colonial} = \frac{\%\partial price}{\partial colonial} = 5.4\%$$

As the dummy for colonial style changes from 0 to 1, the house price increases by 5.4 percentage points

Dummy variables for multiple categories

- **Using dummy variables for multiple categories**
  - 1) Define membership in each category by a dummy variable
  - 2) Leave out one category (which becomes the base category)

$$\widehat{\log}(wage) = \underset{}{.321} + \underset{(.100)}{.213}\, marrmale - \underset{(.058)}{.198}\, marrfem - \underset{(.056)}{.110}\, singfem + \underset{(.007)}{.079}\, educ + \underset{(.005)}{.027}\, exper - \underset{(.00011)}{.00054}\, exper^2 + \underset{(.007)}{.029}\, tenure - \underset{(.00023)}{.00053}\, tenure^2$$

Holding other things fixed, married women earn 19.8% less than single men (= the base category)

Ordinal Dummy Variables

- **Incorporating ordinal information using dummy variables**
- **Example: City credit ratings and municipal bond interest rates**

Municipal bond rate

Credit rating from 0-4 (0=worst, 4=best)

$$MBR = \beta_0 + \beta_1 CR + other\ \ factors$$

This specification would probably not be appropriate as the credit rating only contains ordinal information. A better way to incorporate this information is to define dummies:

$$MBR = \beta_0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 + other\ \ factors$$

Dummies indicating whether the particular rating applies, e.g. CR_1=1 if CR=1 and CR_1=0 otherwise. All effects are measured in comparison to the worst rating (= base category).

Interactions among dummy variables

- **Interactions involving dummy variables**
- **Allowing for different slopes**

Interaction term

$$\log(wage) = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female \cdot educ + u$$

β_0 = intercept men $\quad$ β_1 = slope men

$\beta_0 + \delta_0$ = intercept women $\quad$ $\beta_1 + \delta_1$ = slope women

- **Interesting hypotheses**

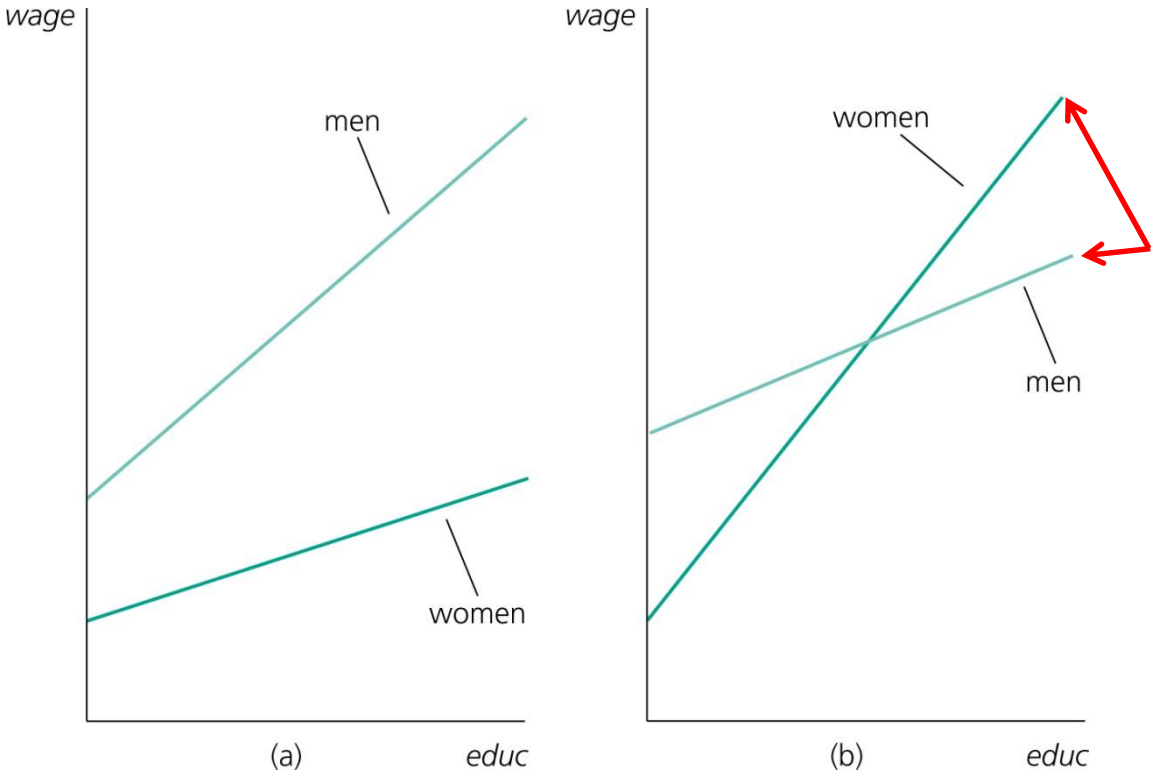
$H_0 : \delta_1 = 0$

The return to education is the same for men and women

$H_0 : \delta_0 = 0, \delta_1 = 0$

The whole wage equation is the same for men and women

Graphical illustration



Interacting both the intercept and the slope with the female dummy enables one to model completely independent wage equations for men and women

Dummy-Continuous /Discrete Interaction (2)

- Testing for differences in regression functions across groups
- Unrestricted model (contains full set of interactions)

College grade point average

Standardized aptitude test score

High school rank percentile

$$cumgpa = \beta_0 + \delta_0 female + \beta_1 sat + \delta_1 female \cdot sat + \beta_2 hsperc$$

$$+ \delta_2 female \cdot hsperc + \beta_3 tothrs + \delta_3 female \cdot tothrs + u$$

Total hours spent in college courses

- Restricted model (same regression for both groups)

$$cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + u$$

Dummy-Continuous /Discrete Interaction (3)

- **Null hypothesis**

$$H_0 : \delta_0 = 0, \delta_1 = 0, \delta_2 = 0, \delta_3 = 0$$

All interaction effects are zero, i.e. the same regression coefficients apply to men and women

- **Estimation of the unrestricted model**

$$\widehat{cumgpa} = \underset{(.21)}{1.48} - \underset{(.411)}{.353}\, female + \underset{(.0002)}{.0011}\, sat + \underset{(.00039)}{.00075}\, female \cdot sat$$

$$- \underset{(.0014)}{.0085}\, hisperc - \underset{(.00316)}{.00055}\, female \cdot hisperc$$

$$+ \underset{(.0009)}{.0023}\, tothrs - \underset{(.00163)}{.00012}\, female \cdot tothours$$

Tested individually, the hypothesis that the interaction effects are zero cannot be rejected

Models (with Dummy Variables)

- **Joint test with F-statistic**

Null hypothesis is rejected

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{(85.515 - 78.355)/4}{78.355/(366-7-1)} \approx 8.18$$

- SSRr is the sum of squared residuals from the restricted regression, i.e., the regression where we impose the restriction.
- SSRur is the sum of squared residuals from the full model,
- q is the number of restrictions under the null and
- k is the number of regressors in the unrestricted regression.

Binary dependent variable

- A Binary dependent variable: the linear probability model
- Linear regression when the dependent variable is binary

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

$$\Rightarrow \quad E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

$$E(y|\mathbf{x}) = 1 \cdot P(y = 1|\mathbf{x}) + 0 \cdot P(y = 0|\mathbf{x})$$

$$\Rightarrow \quad \boxed{P(y = 1|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k}$$

$$\Rightarrow \quad \beta_j = \partial P(y = 1|\mathbf{x})/\partial x_j$$

If the dependent variable only takes on the values 1 and 0

Linear probability model (LPM)

In the linear probability model, the coefficients describe the effect of the explanatory variables on the probability that y=1

Binary dependent variable:Example-1

- **Example: Labor force participation of married women**

=1 if in labor force, =0 otherwise

Non-wife income (in thousand dollars per year)

$$\widehat{inlf} = \underset{(.154)}{.586} - \underset{(.0014)}{.0034}\, nwifeinc + \underset{(.007)}{.038}\, educ + \underset{(.006)}{.039}\, exper$$

$$- \underset{(.00018)}{.00060}\, exper^2 - \underset{(.002)}{.016}\, age - \underset{(.034)}{.262}\, kidslt6$$

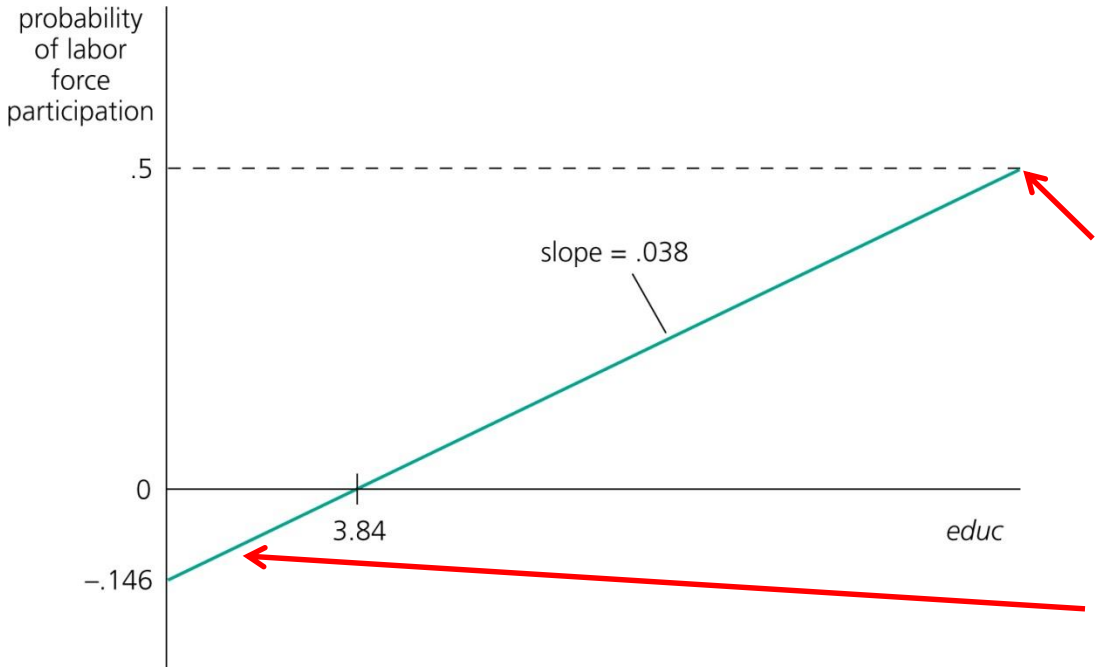
$$+ \underset{(.0132)}{.0130}\, kidsge6, \quad n = 753, R^2 = .264$$

If the number of kids under six years increases by one, the pro- probability that the woman works falls by 26.2%

Does not look significant (but see below)

Binary dependent variable:Example-2

- **Example: Female labor participation of married women (cont.)**



Graph for nwifeinc=50, exper=5, age=30, kindslt6=1, kidsge6=0

The maximum level of education in the sample is educ=17. For the given case, this leads to a predicted probability to be in the labor force of about 50%.

Negative predicted probability *(but no problem because no woman in the sample has educ < 5)*.