

# FINM3123 Introduction to Econometrics

## Chapter 7

### Multiple Regression Analysis with Qualitative Information

# Multiple Regression Analysis: Qualitative Information

## Qualitative Information

- Examples: gender, race, industry, region, rating grade, ...
- A way to incorporate qualitative information is to use **dummy variables** (a.k.a. **binary variables**)
- They may appear as the dependent or as independent variables

## A single dummy independent variable

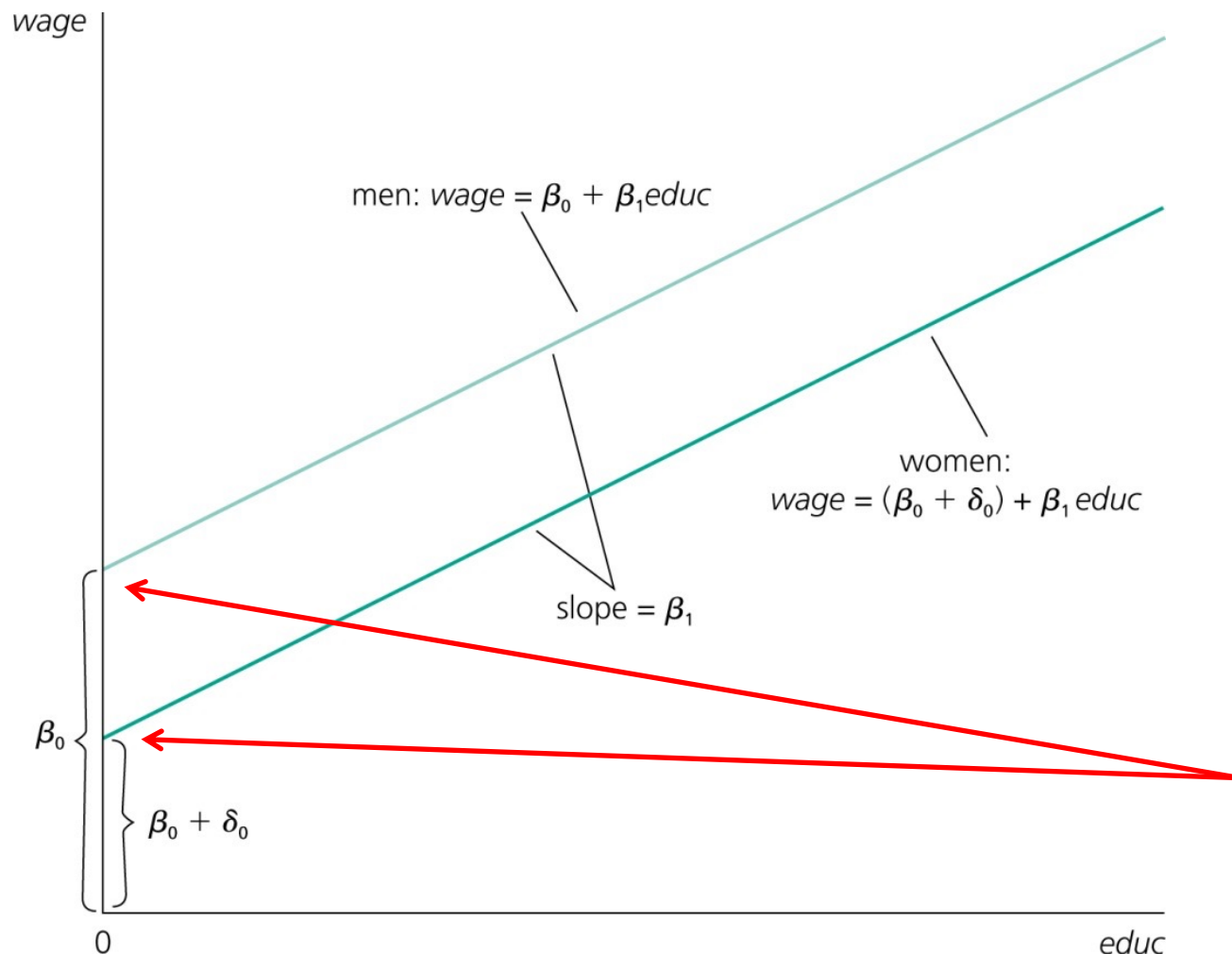
$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u$$

= 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

# Multiple Regression Analysis: Qualitative Information

## 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**

# Multiple Regression Analysis: Qualitative Information

## Dummy variable trap

This model cannot be estimated (perfect collinearity)

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

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

$$wage = \beta_0 + \delta_0 \text{female} + \beta_1 educ + u$$

The base category are men

$$wage = \beta_0 + \gamma_0 \text{male} + \beta_1 educ + u$$

The base category are women

Alternatively, one could omit the intercept:

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

Disadvantages:

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

# Multiple Regression Analysis: Qualitative Information

## Estimated wage equation with intercept shift

$$\widehat{wage} = -1.57_{(.72)} - 1.81_{(.26)} female + .572_{(.049)} educ \\ + .025_{(.012)} exper + .141_{(.021)} tenure$$

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

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

## 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.

# Multiple Regression Analysis: Qualitative Information

## Comparing means of subpopulations described by dummies

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

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

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


- Whether the difference in means is significant can easily be tested
- 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

# Multiple Regression Analysis: Qualitative Information

## Further example: effects of training grants on hours of training

Hours training per employee

Dummy indicating whether firm received training grant


$$\widehat{hrsemp} = 46.67 + 26.25 \text{ grant} - 0.98 \log(sales) - 6.07 \log(employ), \quad n = 105, R^2 = .237$$

(43.41)      (5.59)                      (3.54)                      (3.88)

This is an example of **program evaluation**

- **Treatment group** (= grant receivers) vs. **control group** (= no grant)
- Is the effect of treatment on the outcome of interest causal?

# Multiple Regression Analysis: Qualitative Information

## Using dummy explanatory variables in equations for $\log(y)$

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

(.65)      (.038)                      (.093)

$$+ .027 \text{ } bdrms + .054 \text{ } colonial$$

(.029) (.045)

Dummy variable indicating whether the house is of colonial style

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

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

As the dummy for colonial style changes from 0 to 1, the house price increases by about 5.4%. More precisely, it increases by  $\exp(5.4\%) - 1 = 5.55\%$



# Multiple Regression Analysis: Qualitative Information

## 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**)

$$\begin{aligned}\widehat{\log(wage)} = & .321 + .213 \text{ marrmale} - .198 \text{ marrfem} \\ & (.100) \quad (.055) \quad (.058) \\ & - .110 \text{ singfem} + .079 \text{ educ} + .027 \text{ exper} - .00054 \text{ exper}^2 \\ & (.056) \quad (.007) \quad (.005) \quad (.00011) \\ & + .029 \text{ tenure} - .00053 \text{ tenure}^2 \\ & (.007) \quad (.00023)\end{aligned}$$

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

# Multiple Regression Analysis: Qualitative Information

## Incorporating ordinal information using dummy variables

Example: city credit ratings and municipal bond interest rates

Municipal bond rate


Credit rating from 0 to 4 (0=worst, 4=best): **ordinal variable**



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

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:

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

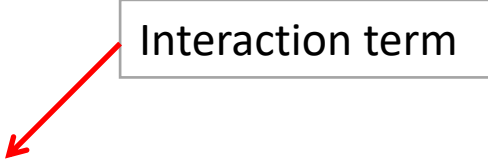


Dummies indicating whether the particular rating applies, e.g.  $CR_1 = 1$  if  $CR = 1$  and  $CR_1 = 0$  otherwise. All effects are measured in comparison to the worst rating 0 (= base category).

# Multiple Regression Analysis: Qualitative Information

## Interactions involving dummy variables


- Allowing for different slopes

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



$$\beta_0 = \text{intercept men} \qquad \beta_1 = \text{slope men}$$

$$\beta_0 + \delta_0 = \text{intercept women} \qquad \beta_1 + \delta_1 = \text{slope women}$$

- Interesting hypotheses:

$$\boxed{H_0 : \delta_1 = 0}$$


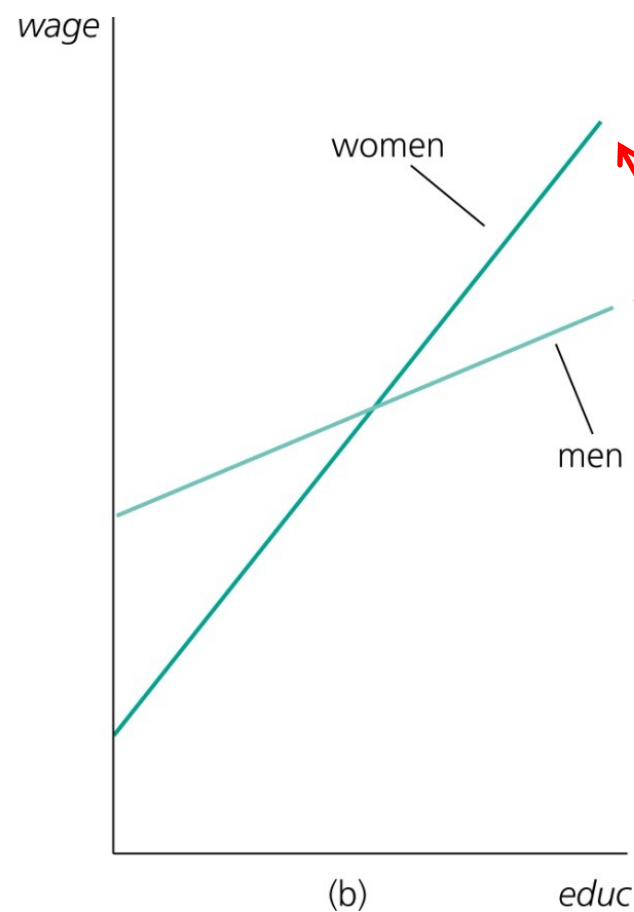
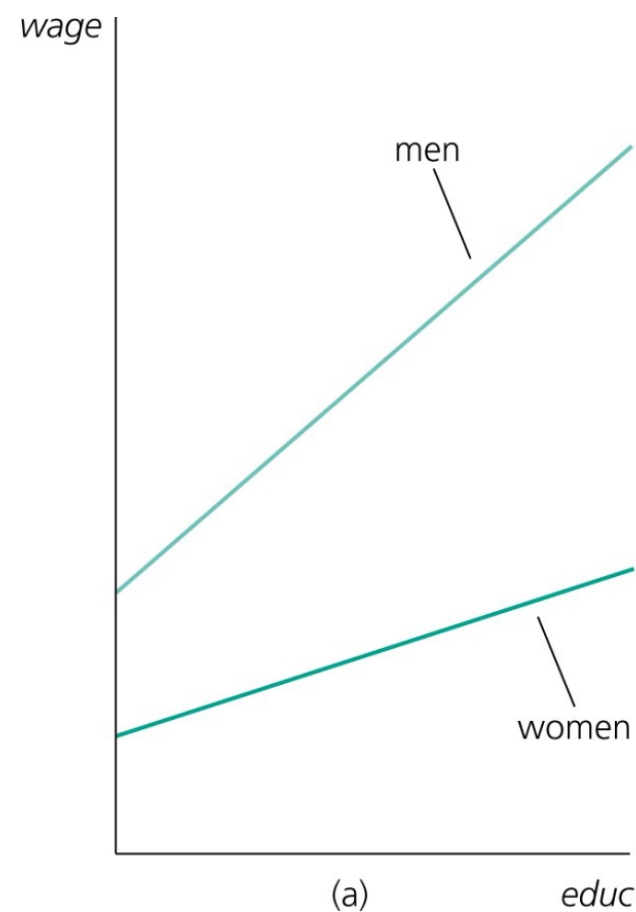
The return to education is the same for men and women

$$\boxed{H_0 : \delta_0 = 0, \delta_1 = 0}$$


The whole wage equation is the same for men and women

# Multiple Regression Analysis: Qualitative Information

## 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

# Multiple Regression Analysis: Qualitative Information

## Estimated wage equation with interaction term

$$\begin{aligned}\widehat{\log(wage)} = & .389 - .227 \text{ female} - .082 \text{ educ} \\ & (.119) \quad (.168) \quad (.008) \\ & - .0056 \text{ female} \cdot \text{educ} + .029 \text{ exper} - .00058 \text{ exper}^2 \\ & (.0131) \quad (.005) \quad (.00011) \\ & + .032 \text{ tenure} - .00059 \text{ tenure}^2, n = 526, R^2 = .441 \\ & (.007) \quad (.00024)\end{aligned}$$

No evidence against hypothesis that the return to education is the same for men and women

Does this mean that there is no significant evidence of lower pay for women at the same levels of educ, exper, and tenure? No: this is only the effect for educ = 0, which is difficult to estimate because very few people in the sample have very low levels of education. To better answer the question one could recenter the interaction term, e.g. around educ = 12.5 (= average education).

# Multiple Regression Analysis: Qualitative Information

## 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

$$\begin{aligned} cumgpa = & \beta_0 + \delta_0 female + \beta_1 sat + \delta_1 female \cdot sat + \beta_2 hsperc \\ & + \delta_2 female \cdot hsperc + \beta_3 tothrs + \delta_3 female \cdot tothrs + u \end{aligned}$$

- Restricted model (same regression for both groups)

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

Total hours spent  
in college courses

# Multiple Regression Analysis: Qualitative Information

## 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

$$\begin{aligned} \widehat{cumgpa} = & 1.48 - .353 \text{ female} + .0011 \text{ sat} + .00075 \text{ female} \cdot \text{sat} \\ & (.21) \quad (.411) \quad (.0002) \quad (.00039) \\ & - .0085 \text{ hisperc} - .00055 \text{ female} \cdot \text{hisperc} \\ & (.0014) \quad (.00316) \\ & + .0023 \text{ tothrs} - .00012 \text{ female} \cdot \text{tothours} \\ & (.0009) \quad (.00163) \end{aligned}$$

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

# Multiple Regression Analysis: Qualitative Information

## Joint test with F-statistic

Null hypothesis is rejected

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

## Alternative way to compute the F-statistic in this given case

- Run separate regressions for men and for women; the unrestricted SSR is given by the sum of the SSR of these two regressions
- Run regression for the restricted model and store SSR
- If the test is computed in this way it is called **Chow test**
- Important: Test assumes a constant error variance across groups



# Multiple Regression Analysis: Qualitative Information

## A binary dependent variable: the linear probability model

- Linear regression when the dependent variable is binary

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

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

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

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

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

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  $y = 1$**

# Multiple Regression Analysis: Qualitative Information

## Example: Labor force participation of married women

=1 if in labor force, =0 otherwise

Non-wife income (in thousand dollars per year)

$$\widehat{inlf} = .586 - .0034 \text{ nwifeinc} + .038 \text{ educ} + .039 \text{ exper}$$

(.154)    (.0014)                    (.007)                    (.006)

$$- .00060 \text{ exper}^2 - .016 \text{ age} - .262 \text{ kidslt6}$$

(.00018)                    (.002)                    (.034)

$$+ .0130 \text{ kidsge6}, n = 753, R^2 = .264$$

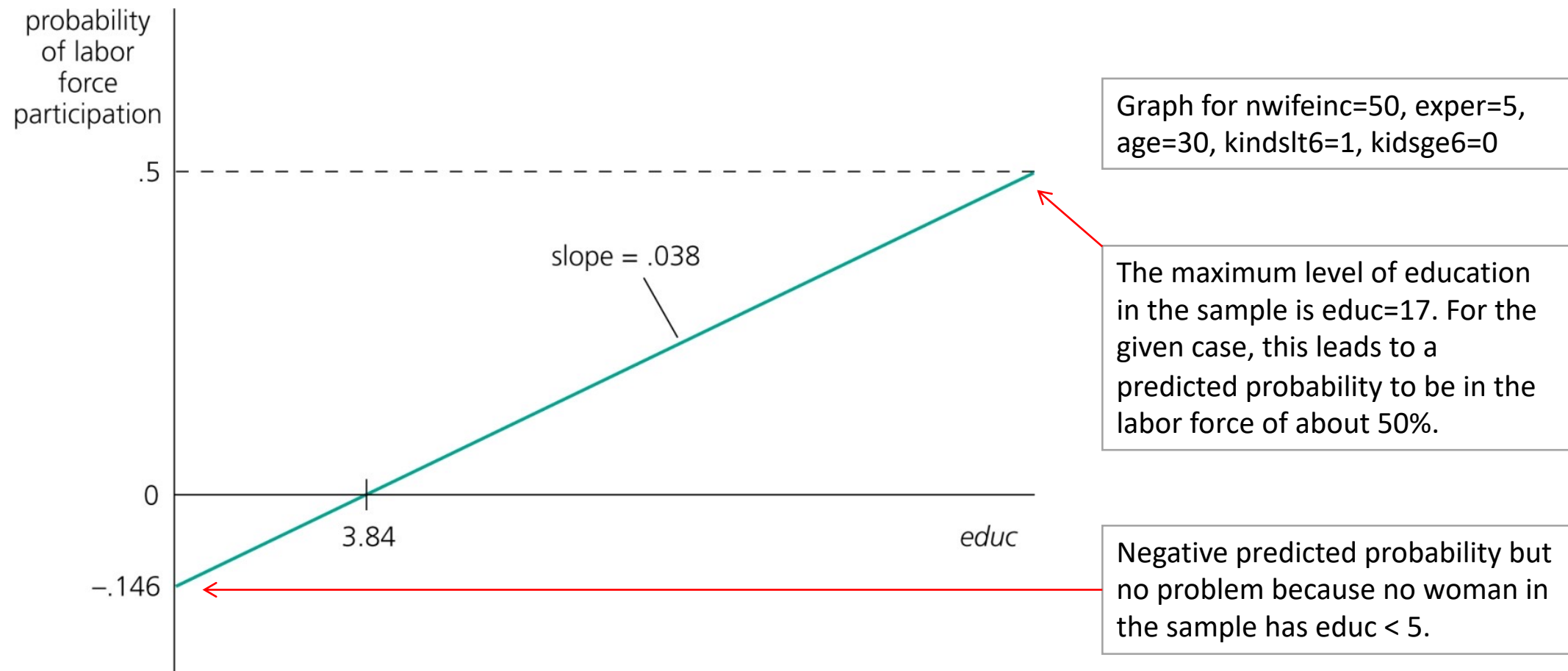
(.0132)

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

Does not look significant

# Multiple Regression Analysis: Qualitative Information

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



# Multiple Regression Analysis: Qualitative Information

## Disadvantages of the linear probability model

- Predicted probabilities may be larger than one or smaller than zero
- Marginal probability effects sometimes logically impossible
- The linear probability model is necessarily heteroskedastic

$$Var(y|\mathbf{x}) = P(y = 1|\mathbf{x}) [1 - P(y = 1|\mathbf{x})] \quad \leftarrow \text{Variance of Bernoulli variable}$$

- Heteroskedasticity consistent standard errors need to be computed

## Advantages of the linear probability model

- Easy estimation and interpretation
- Estimated effects and predictions often reasonably good in practice

# Multiple Regression Analysis: Qualitative Information

## More on policy analysis and program evaluation

Example: Effect of job training grants on worker productivity

Percentage of defective items

=1 if firm received training grant, =0 otherwise

$$\widehat{\log(scrap)} = 4.99 - .052 \text{ grant} - .455 \log(sales)$$

(4.66) (.431) (.373)

$$+ .639 \log(employ), n = 50, R^2 = .072$$

(.365)

No apparent effect of grant on productivity

**Treatment group:** grant receivers, **Control group:** firms that received no grant

Grants were given on a first-come, first-served basis. This is not the same as giving them out randomly. It might be the case that firms with less productive workers saw an opportunity to improve productivity and applied first.

# Multiple Regression Analysis: Qualitative Information

## Self-selection into treatment as a source for endogeneity

- In the given and in related examples, the treatment status is probably related to other characteristics that also influence the outcome
- The reason is that subjects self-select themselves into treatment depending on their individual characteristics and prospects

## Experimental evaluation

- In experiments, assignment to treatment is random
- In this case, causal effects can be inferred using a simple regression

$$y = \beta_0 + \beta_1 \text{partic} + u$$

The dummy indicating whether or not there was treatment is unrelated to other factors affecting the outcome.

# Multiple Regression Analysis: Qualitative Information

## Further example of an endogenous dummy regressor

- Are nonwhite customers discriminated against?

$$\text{approved} = \beta_0 + \beta_1 \text{nonwhite} + \beta_2 \text{income} + \beta_3 \text{wealth} + \beta_4 \text{credrate} + u$$

- It is important to control for other characteristics that may be important for loan approval (e.g. profession, unemployment)
- Omitting important characteristics that are correlated with the non-white dummy will produce spurious evidence for discrimination

# Summary

- Interpretation of coefficients of dummy variables
- Dummy variable trap
- Changing base variable to make statistical inference easier
- Incorporating ordinal information using dummy variables
  - Include multiple 0-1 dummy variables instead of one variable taking multiple ordinal values



# Summary

- Interactions involving dummy variables
  - Interpretation of the coefficient of the interaction term
  - Hypothesis testing about group differences
  - Chow-Test
- Linear probability model
  - Interpretation of intercept and slope coefficients
  - Heteroskedasticity issue