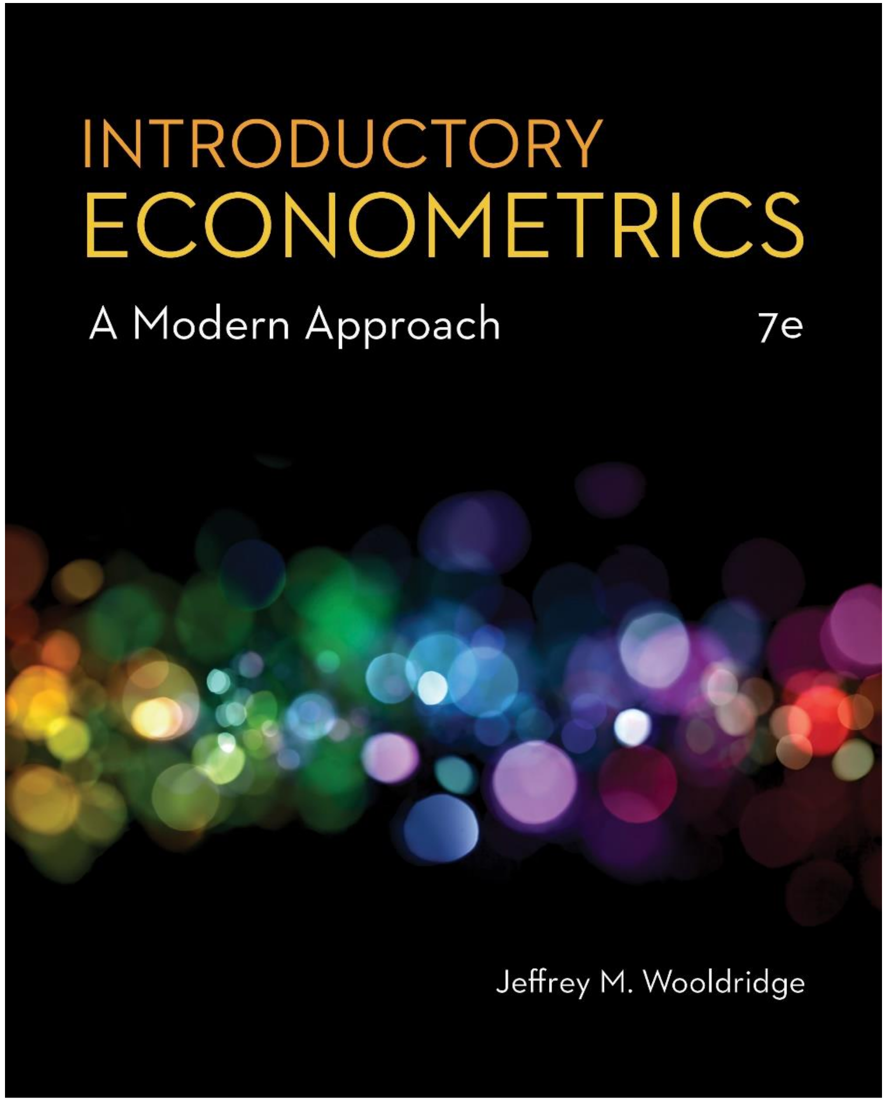
Chapter 7

Multiple Regression Analysis with Qualitative Information



Multiple Regression Analysis with Qualitative Information (1 of 24)

- **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**

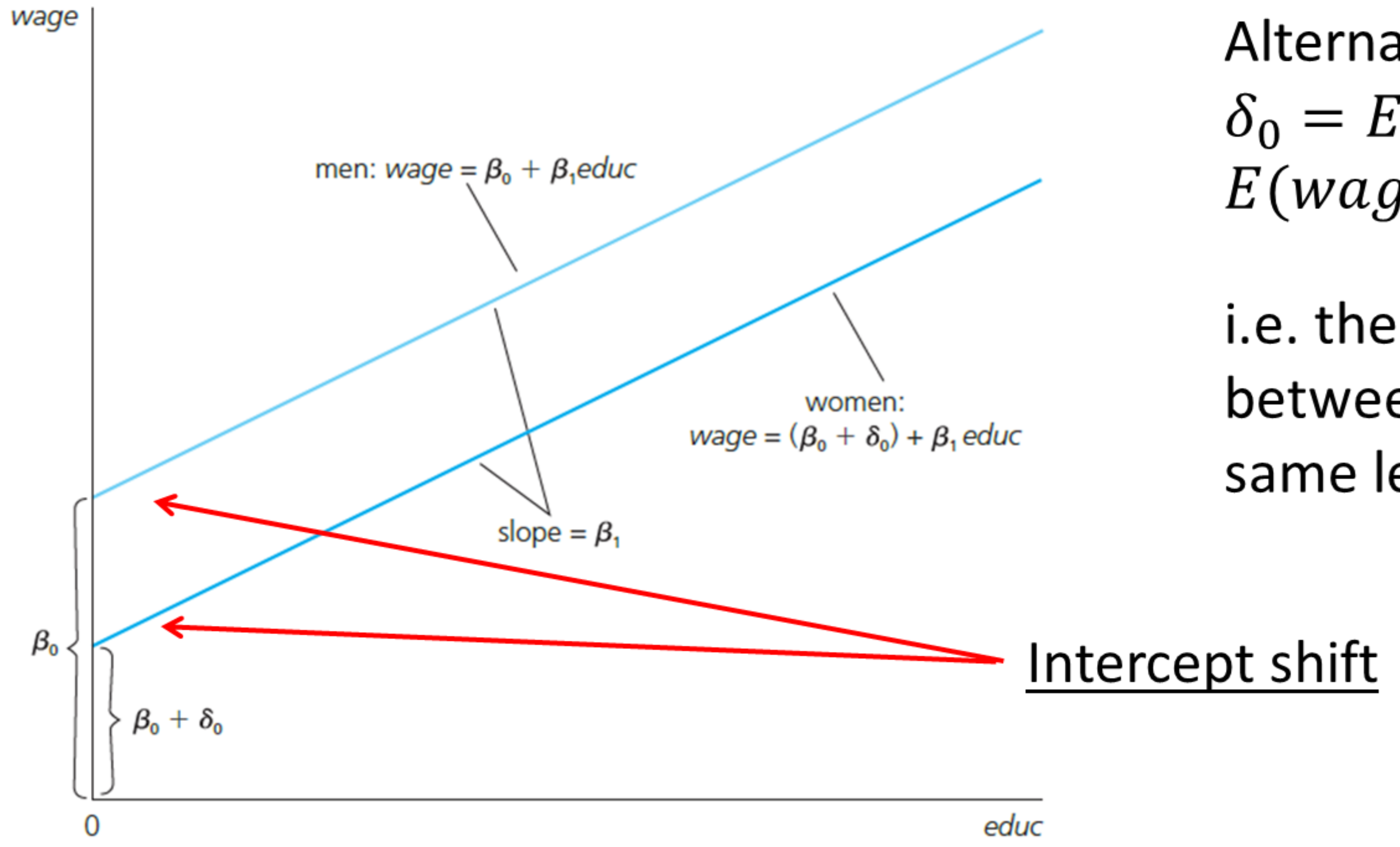
$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u$$

= 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 a man

Multiple Regression Analysis with Qualitative Information (2 of 24)

- 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.

Multiple Regression Analysis with Qualitative Information (3 of 24)

- **Dummy variable trap**

This model cannot be estimated due to perfect collinearity

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

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

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

The base category are men

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

The base category are women

- Alternatively, one could omit the intercept

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

Disadvantages:

1. More difficult to test for differences between the parameters.
2. R-squared formula invalid without an intercept

Multiple Regression Analysis with Qualitative Information (4 of 24)

- **Estimated wage equation with intercept shift**

WAGE1.dta

$$\widehat{wage} = \underset{(0.72)}{-1.57} - \underset{(0.26)}{1.81}\,female + \underset{(0.049)}{0.572}\,educ$$
$$+\underset{(0.012)}{0.025}\,exper + \underset{(0.021)}{0.141}\,tenure$$

$$n = 526, R^2 = 0.364$$

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

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

Multiple Regression Analysis with Qualitative Information (5 of 24)

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

$$\widehat{wage} = \underset{(0.21)}{7.10} - \underset{(0.26)}{2.51}\, female$$

$$n = 526, R^2 = 0.116$$

Not holding other factors constant, women earn $2.51 less than men; i.e. the difference between the mean wages of men and women is $2.51

- **Discussion**
  - It can easily be tested whether the 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.

Multiple Regression Analysis with Qualitative Information (6 of 24)

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

Hours training per employee

Dummy variable indicating whether firm received a training grant

JTRAIN.dta

$$\widehat{hrsemp} = \underset{(43.41)}{46.67} + \underset{(5.59)}{26.25}\,grant - \underset{(3.54)}{0.98}\,sales - \underset{(3.88)}{6.07}\log(employ)$$

$$n = 105, R^2 = 0.237$$

- **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? Counterfactuals!

Multiple Regression Analysis with Qualitative Information (7 of 24)

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

$$\widehat{\log}(price) = \underset{(0.65)}{-1.35} + \underset{(0.038)}{0.168}\log(lotsize) + \underset{(0.093)}{0.707}\log(sqrft)$$
$$+\underset{(0.029)}{0.027}bdrms + \underset{(0.045)}{0.054}colonial$$
$$n = 88, R^2 = 0.649$$

Dummy indicating whether house is of colonial style

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

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

Multiple Regression Analysis with Qualitative Information (8 of 24)

- **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)

WAGE1.dta

$$\widehat{\log(wage)} = \underset{(0.100)}{0.321} + \underset{(0.055)}{0.213}marrmale - \underset{(0.058)}{0.198}marrfem$$
$$-\underset{(0.056)}{0.110}singfem + \underset{(0.007)}{0.079}educ + \underset{(0.005)}{0.027}exper - \underset{(0.00023)}{0.00054}exper^2$$
$$+\underset{(0.007)}{0.079}tenure - \underset{(0.00023)}{0.00053}tenure^2$$

$$n = 2{,}725, R^2 = 0.0422$$

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

i.married#i.female

Multiple Regression Analysis with Qualitative Information (9 of 24)

- **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)

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

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).

Multiple Regression Analysis with Qualitative Information (10 of 24)

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

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

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

δ_0 = intercept for women $\quad$ δ_1 = slope for women

Interaction term

- **Interesting hypothesis**

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

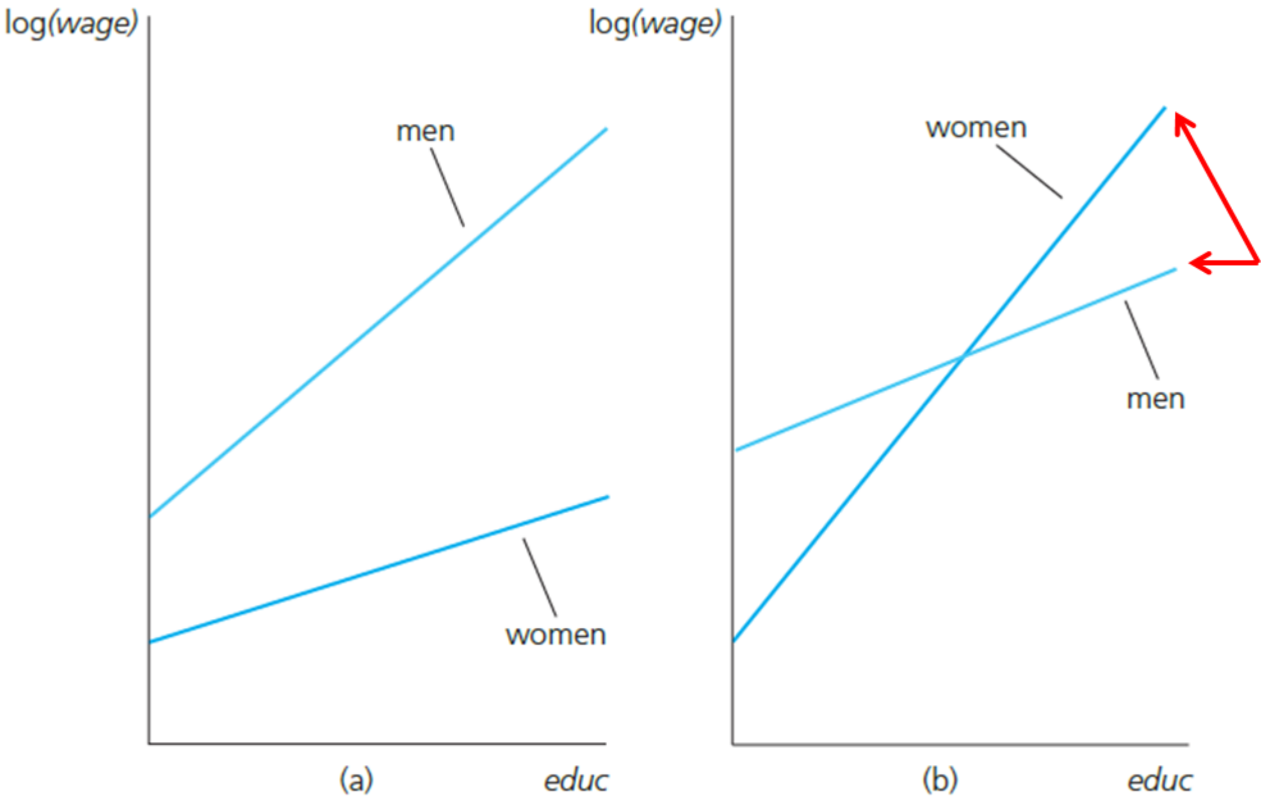
The return to education is the same for men and women

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

The whole wage equation is the same for men and women

Multiple Regression Analysis with Qualitative Information (11 of 24)

- **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 with Qualitative Information (12 of 24)

- **Estimated wage equation with interaction term**

$$\widehat{\log}(wage) = \underset{(.119)}{.389} - \boxed{\underset{(.168)}{.227}}\ female - \underset{(.008)}{.082}\ educ$$

$$- \boxed{\underset{(.0131)}{.0056}}\ female{\cdot}educ + \underset{(.005)}{.029}\ exper - \underset{(.00011)}{.00058}\ exper^2$$

$$+ \underset{(.007)}{.032}\ tenure - \underset{(.00024)}{.00059}\ tenure^2, n = 526, R^2 = .441$$

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? <u>No: this is only the effect for educ = 0.</u> To answer the question one has to recenter the interaction term, e.g. around educ = 12.5 (= average education).

Multiple Regression Analysis with Qualitative Information (13 of 24)

- **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)**

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

Multiple Regression Analysis with Qualitative Information (14 of 24)

- **Null hypothesis**

$$\boxed{H_0 : \delta_0 = 0, \delta_1 = 0, \delta_2 = 0, \delta_3 = 0}$$

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

GPA3.dta

- **Estimation of the unrestricted model**

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

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

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

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

$$n = 366, R^2 = .406, \overline{R^2} = .394$$

Multiple Regression Analysis with Qualitative Information (15 of 24)

- **Joint test with F-statistic**

Null hypothesis is rejected

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

- **Alternative way to compute F-statistic in the 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 the Chow-Test.
  - Important: Test assumes a constant error variance across groups.

Multiple Regression Analysis with Qualitative Information (16 of 24)

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

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

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

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

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

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

<u>Linear probability model (LPM)</u>

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

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

Multiple Regression Analysis with Qualitative Information (17 of 24)

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

MROZ.dta

=1 if in labor force, =0 otherwise

Non-wife income (in thousand dollars per year)

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

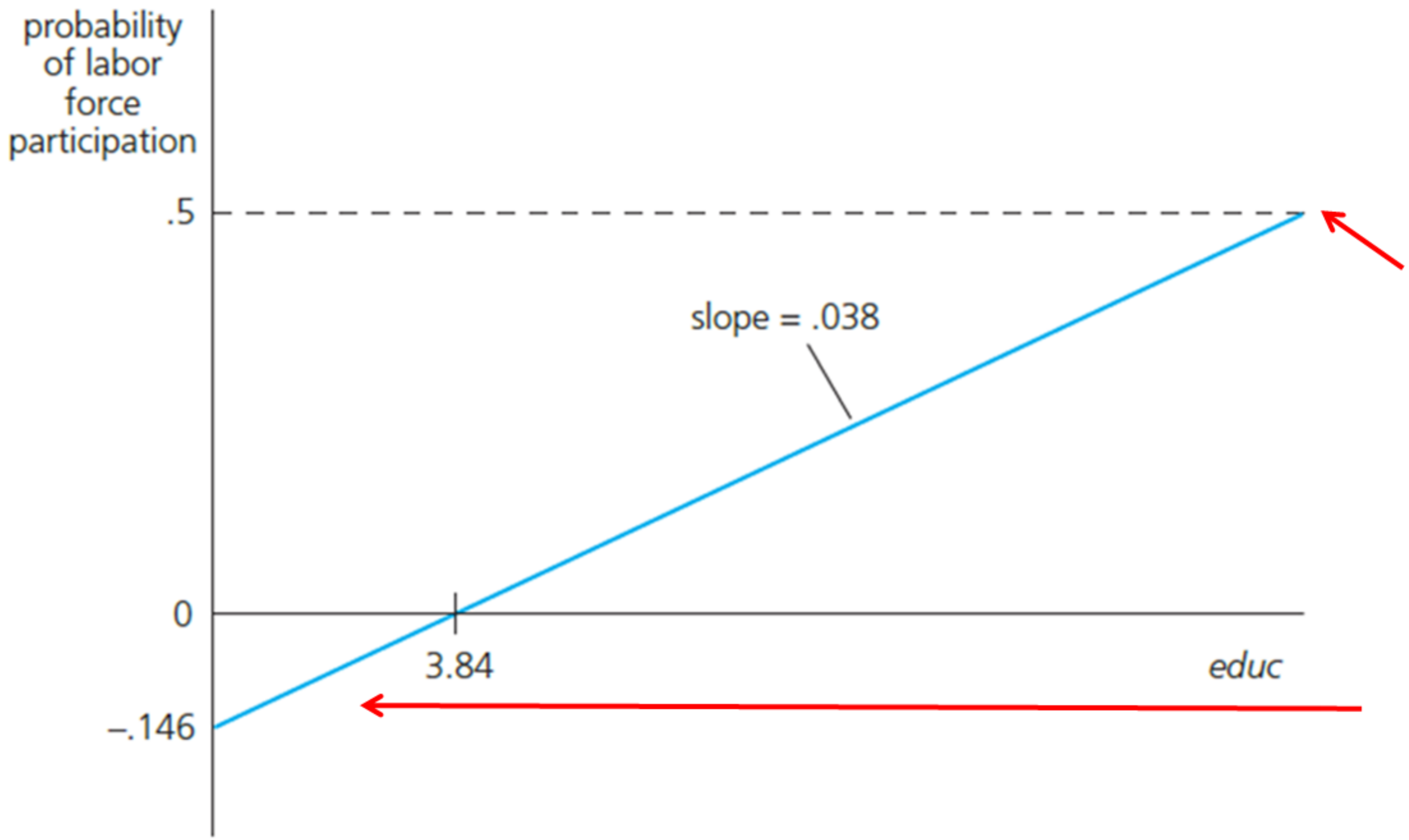
$$- \underset{(.00018)}{.00060}\ exper^2 - \underset{(.002)}{.016}\ age - \underset{(.034)}{.262}\ kidslt6$$

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

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

Multiple Regression Analysis with Qualitative Information (18 of 24)

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



Graph for nwifeinc=50, exper=5, age=30, kindslt6=1, and 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%.

There is a negative predicted probability, but no problem because no woman in the sample has educ < 5.

Multiple Regression Analysis with Qualitative Information (19 of 24)

- **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.
  - Thus, heteroskedasticity consistent standard errors need to be computed.

$$Var(y|\mathrm{x}) = P(y=1|\mathrm{x})\,[1 - P(y=1|\mathrm{x})]$$

Variance of Bernoulli variable

- **Advantages of the linear probability model**
  - Easy estimation and interpretation
  - Estimated effects and predictions are often reasonably good in practice.

Multiple Regression Analysis with Qualitative Information (20 of 24)

- **More on policy analysis and program evaluation**
- Example: Effect of job training grants on worker productivity

The firm's scrap rate

=1 if firm received training grant, =0 otherwise

$$\widehat{\log}(scrap) = \underset{(4.66)}{4.99} \boxed{\underset{(.431)}{-\ .052}}\ grant - \underset{(.373)}{.455}\ \log(sales)$$

No apparent effect of grant on productivity

$$+ \underset{(.365)}{.639}\ \log(employ), n = 50, R^2 = .072$$

- 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 with Qualitative Information (21 of 24)

- **Addressing the problem of self-selection**

$$E(y|w,x) = \alpha + \tau w + \gamma_1 x_1 + \cdots + \gamma_k x_k$$
$$\rightarrow y = (1-w)y(0) + wy(1)$$

$y(0)$ is the outcome of y when $w = 0$
$y(1)$ is the outcome of y when $w = 1$

w is a treatment indicator equal to 1 when the treatment has been applied

- We include x_1 through x_j to account for the possibility that the treatment (w) is not randomly assigned.
- For example, children eligible for a program like Head Start participate based on parental decisions. We thus need to control for things like family background and structure to get closer to random assignment into the treatment (participates in Head Start) and control (does not participate) groups.

Multiple Regression Analysis with Qualitative Information (22 of 24)

- **Addressing the problem of self-selection continued (cont.)**
  - Consider the simple regression:

$$y = \alpha + \tau w + u$$

- We need to make the strong assumption that w is independent of [y(0),y(1)]. In other words, treatment is randomly assigned.
  - A more convincing case is to include covariates x_1 through x_j

$$y = \alpha + \tau w + \gamma_1 x_1 + \cdots + \gamma_j x_j + u$$

The estimator $\hat{\tau}$ is the regression adjusted estimator

- Now we assume that w is independent of [y(0),y(1)] conditional upon x_1 through x_j.
  - This is known as **regression adjustment** and allows us to adjust for differences across units in estimating the causal effect of the treatment.

Multiple Regression Analysis with Qualitative Information (23 of 24)

- **Relaxing the assumption of a constant treatment effect**
  - We can allow the treatment effect to vary across observations and instead estimate the **average treatment effect** (ATE)

$$y_i = \alpha + \tau w_i + \gamma_1 x_{i1} + \cdots + \gamma_k x_{ik} + \delta_1 w_i(x_{i1} - \bar{x}_1) + \cdots + \delta_k w_i(x_{ik} - \bar{x}_k) + u$$

$\bar{x}_1, \dots, \bar{x}_k$ are the sample averages of x_{i1} through x_{ik}

- The estimated coefficient on w will be the ATE.
- The regression that allows individual treatment effects to vary is known as the **unrestricted regression adjustment** (URA).
- By contrast, a **restricted regression adjustment** (RRA) forces the treatment effect to be identical across individuals.

Multiple Regression Analysis with Qualitative Information (24 of 24)

- **An alternative method for obtaining the URA ATE**

Control: $\hat{y}_i^{(0)} = \hat{\alpha} + \hat{\gamma}_{0,1}x_{i,1} + \cdots + \hat{\gamma}_{0,k}x_{i,k}$ using n_0 control observations

Treatment: $\hat{y}_i^{(1)} = \hat{\alpha} + \hat{\gamma}_{1,1}x_{i,1} + \cdots + \hat{\gamma}_{1,k}x_{i,k}$ using n_1 control observations

- Now for every unit in the sample, predict $y_i(0)$ and $y_i(1)$ regardless of whether the unit is in the control or treatment groups.
  - Use these predicted values to compute the ATE as:

$$\frac{1}{n}\sum_{i=1}^{n}[\hat{y}_i^{(1)} - \hat{y}_i^{(0)}]$$

- Though this yields the same ATE as running the regression with interaction terms, computing a standard error by hand can be tricky.