Multiple Regression Analysis with Qualitative Information

Chapter 6 (Ch. 7 of the textbook)

Wooldridge: Introductory Econometrics:
A Modern Approach, 5e

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
  - They may appear as the dependent or as independent variables
- **A single dummy independent variable**

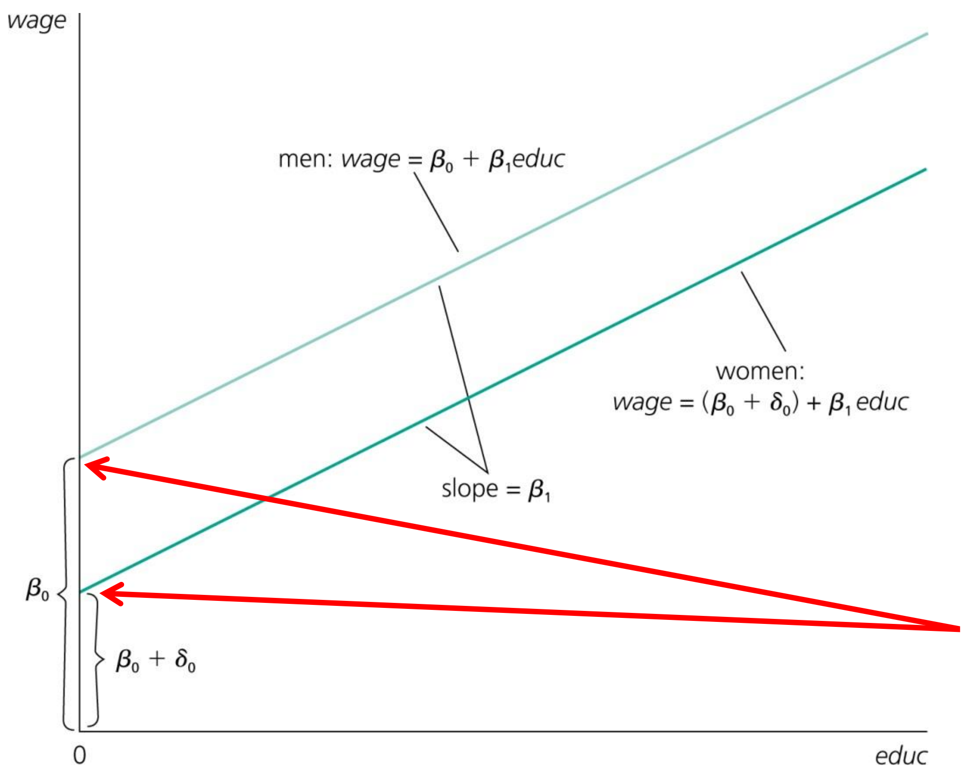
$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u$$

δ_0 = the wage gain/loss if the person is a woman rather than a man (holding other things fixed)

$female$: 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)

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

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:

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

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

Multiple Regression Analysis: Qualitative Information

- **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

- **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**

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

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

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 differ-ences 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 → $\widehat{hrsemp}$

Dummy indicating whether firm received training grant → $grant$

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

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

- **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>?

Multiple Regression Analysis: Qualitative Information

- **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%

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)

$$\widehat{\log}(wage) = \underset{(.100)}{.321} + \underset{(.055)}{.213}\, marrmale - \underset{(.058)}{.198}\, marrfem$$

$$- \underset{(.056)}{.110}\, singfem + \underset{(.007)}{.079}\, educ + \underset{(.005)}{.027}\, exper - \underset{(.00011)}{.00054}\, exper^2$$

$$+ \underset{(.007)}{.029}\, tenure - \underset{(.00023)}{.00053}\, tenure^2$$

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

Multiple Regression Analysis: Qualitative Information

- **Using dummy variables for multiple categories**
- **Exact variations:** $(e^{0.213}-1)\times 100\% = 23.74\%$

$$(e^{-0.198}-1)\times 100\% = -17.96\% \quad (e^{-0.11}-1)\times 100\% = -10.42\%$$

$$\widehat{\log}(wage) = \underset{(.100)}{.321} + \underset{(.055)}{.213}\ marrmale - \underset{(.058)}{.198}\ marrfem$$

$$- \underset{(.056)}{.110}\ singfem + \underset{(.007)}{.079}\ educ + \underset{(.005)}{.027}\ exper - \underset{(.00011)}{.00054}\ exper^2$$

$$+ \underset{(.007)}{.029}\ tenure - \underset{(.00023)}{.00053}\ tenure^2$$

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

Multiple Regression Analysis: Qualitative Information

- **Using dummy variables for multiple categories**

Changing the base category

$$\log(wage) = \underset{}{0.123} + \underset{(0.046)}{0.411}\, marrmale + \underset{(0.058)}{0.198}\, singmale + \underset{(0.052)}{0.088}\, singfemale$$

$$+ 0.079\, educ + \ldots$$

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

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 0-4 (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: Qualitative Information

- **<u>Interactions involving dummy variables</u>**
- **Allowing for different slopes**

Interaction term

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

β_0 = <u>intercept men</u> β_1 = <u>slope men</u>

$\beta_0 + \delta_0$ = <u>intercept women</u> $\beta_1 + \delta_1$ = <u>slope women</u>

- **Interesting hypotheses**

$H_0 : \delta_1 = 0$

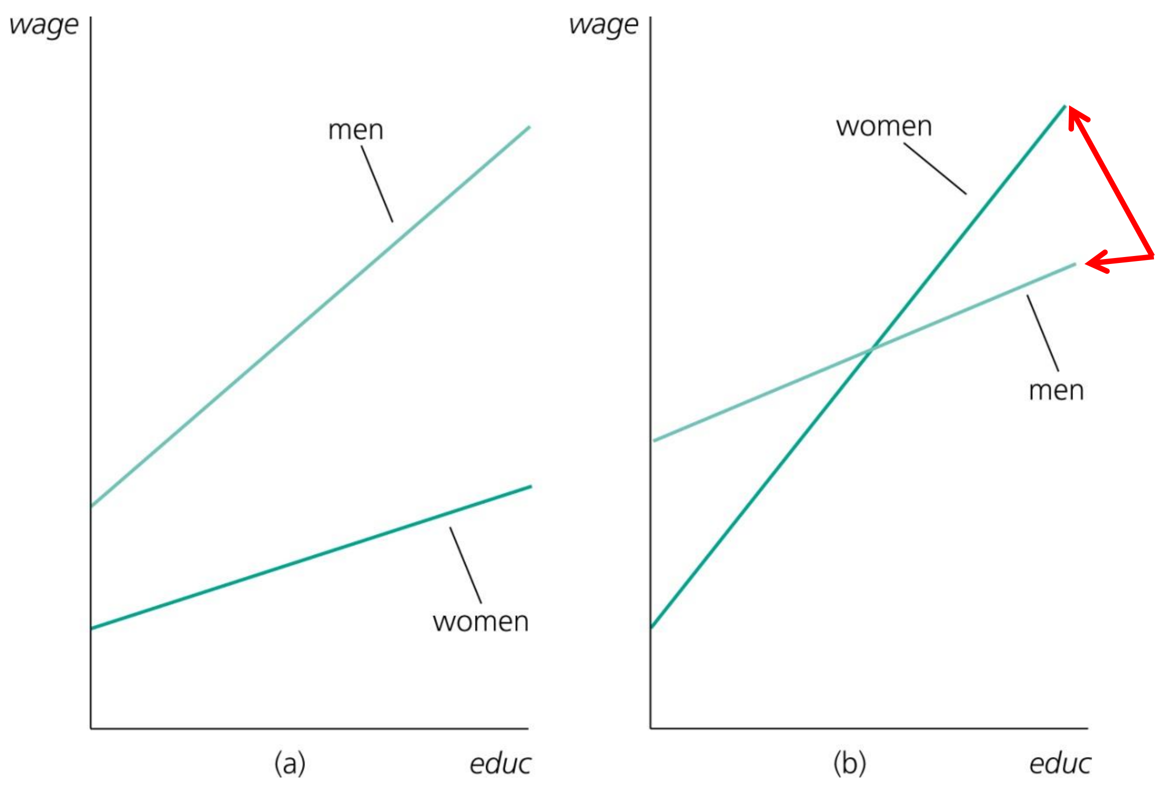
The <u>return to education</u> is the same for men and women

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

The <u>whole wage equation</u> 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**

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

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

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

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

Multiple Regression Analysis: Qualitative Information

- **<u>Testing for differences in regression functions across groups</u>**
- **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: 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**

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

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 F-statistic**
  - Run separate regressions for men and for women; the unrestricted SSR is given by the sum of the SSR of these two regressions: $SSR_1 + SSR_2$
  - Run regression for the restricted model and store SSR: SSR_P
  - If the test is computed in this way it is called the Chow-Test
  - Important: Test assumes a constant error variance accross groups

Multiple Regression Analysis: Qualitative Information

- **Alternative way to compute F-statistic**

<u>CHOW test</u>: $H_0: \beta_{01} = \beta_{02}, \beta_{11} = \beta_{12}, \beta_{21} = \beta_{22}, \ldots, \beta_{k1} = \beta_{k2}$

$$F = \frac{SSR_P - (SSR_1 + SSR_2)}{SSR_1 + SSR_2} \times \frac{n - 2(k+1)}{k+1} \sim F(k+1, n-2(k+1))$$

If the <u>restricted regression contains a dummy for an intercept shift</u> the statistic to test <u>changes only in the slope coefficients</u> is:

$H_0: \beta_{11} = \beta_{12}, \beta_{21} = \beta_{22}, \ldots, \beta_{k1} = \beta_{k2}$

$$F = \frac{SSR_P^* - (SSR_1 + SSR_2)}{SSR_1 + SSR_2} \times \frac{n - 2(k+1)}{k} \sim F(k, n-2(k+1))$$

With SSR_P^* the sum of squared residuals of the restricted regression

Multiple Regression Analysis: Qualitative Information

- **Example**

Dependent Variable: LOG(WAGE)
Method: Least Squares
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.127998	0.105932	1.208296	0.2275
EDUC	0.090366	0.007468	12.10041	0.0000
EXPER	0.041009	0.005197	7.891606	0.0000
EXPER^2	-0.000714	0.000116	-6.163888	0.0000
Sum squared resid	103.7904			

Multiple Regression Analysis: Qualitative Information

- **Example** Dependent Variable: LOG(WAGE)
  Method: Least Squares
  Sample: **IF FEMALE=1**
  Included observations**: 252**

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.266084	0.141535	1.879985	0.0613
EDUC	0.079195	0.010369	7.637838	0.0000
EXPER	0.022372	0.006664	3.356971	0.0009
EXPER^2	-0.000423	0.000148	-2.853701	0.0047

Sum squared resid 38.38393

Multiple Regression Analysis: Qualitative Information

- **Example**

Dependent Variable: LOG(WAGE)
Method: Least Squares
Sample: **IF FEMALE=0**
Included observations: **274**

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.157291	0.136403	1.153132	0.2499
EDUC	0.090354	0.009267	9.750477	0.0000
EXPER	0.054017	0.006743	8.011343	0.0000
EXPER^2	-0.000914	0.000150	-6.079015	0.0000

Sum squared resid 47.35130

Multiple Regression Analysis: Qualitative Information

- **Example**

Dependent Variable: LOG(WAGE)
Method: Least Squares
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.390483	0.102210	3.820413	0.0001
FEMALE	-0.337187	0.036321	-9.283424	0.0000
EDUC	0.084136	0.006957	12.09407	0.0000
EXPER	0.038910	0.004824	8.066683	0.0000
EXPER^2	-0.000686	0.000107	-6.388842	0.0000
Sum squared resid	89.05862			

Multiple Regression Analysis: Qualitative Information

The CHOW test for the example:

$$H_0 : \beta_{01} = \beta_{02}, \beta_{11} = \beta_{12}, \beta_{21} = \beta_{22}, \beta_{31} = \beta_{32}$$

$$F = \frac{103.7904 - (38.3839 + 47.3513)}{38.3839 + 47.3513} \times \frac{526 - 2(3+1)}{3+1} = 27.272 > F_{0.05}(4, 518) = 2.39$$

To test changes only in the slope coefficients in the example:

$$H_0 : \beta_{11} = \beta_{12}, \beta_{21} = \beta_{22}, \beta_{31} = \beta_{32}$$

$$F = \frac{89.0586 - (38.3839 + 47.3513)}{38.3839 + 47.3513} \times \frac{526 - 2(3+1)}{3} = 6.6931 > F_{0.05}(3, 518) = 2.62$$