Class 21 – 22 MRM: Qualitative Regressors (Part I)

Pedram Jahangiry

Jon M. Huntsman School of Business
Utah State University

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

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

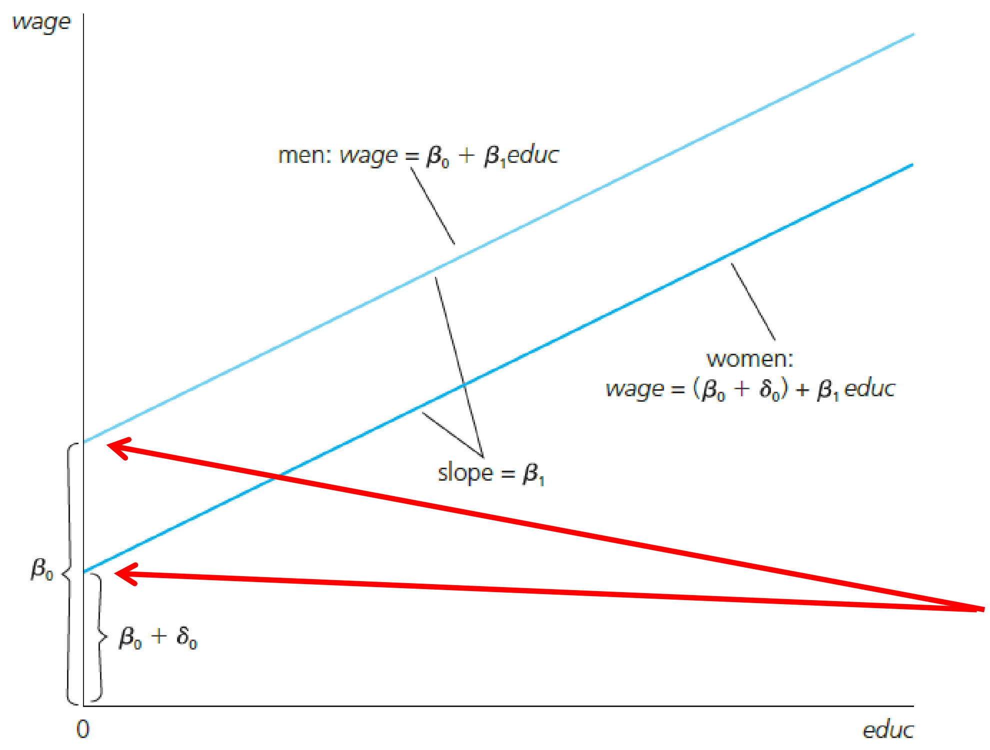
= the wage gain/loss if the person is a woman rather than a man (holding other things fixed)

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

TABLE 7.1 A Partial Listing of the Data in WAGE1

person	*wage*	*educ*	*exper*	*female*	*married*
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

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

Dummy variable trap

This model cannot be estimated (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 (benchmark)** group are men

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

The **base (benchmark)** group are women

Wage Discrimination?

$$wage = \beta_0 + \delta_0 female$$

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

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

Not controling for other factors, on average women earn \$2.51per hour less than men, i.e. the difference between the mean wage of men and that of women is \$2.51.

δ_0 is clearly significant, but 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.

- ❑ What is the average wage for women/men in this example?
- ❑ 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

- ❑ It can easily be tested whether difference in means is significant:

Generally, simple regression on a constant and a dummy variable is a straightforward way to compare the means of two groups.

Wage Discrimination?

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$$

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

$$\widehat{wage} = \underset{(.72)}{-1.57} - \underset{(.26)}{1.81} \, female + \underset{(.049)}{.572} \, educ + \underset{(.012)}{0.25} \, exper + \underset{(.021)}{.141} \, tenure$$

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

δ_0 is still clearly significant, does that mean that women are discriminated against now?

Further example: Effects of computer ownership on college GPA

$$colGPA = \beta_0 + \delta_0 PC + \beta_1 hsGPA + \beta_2 ACT + u,$$

$$\widehat{colGPA} = \underset{(.33)}{1.26} + \underset{(.057)}{.157}\, PC + \underset{(.094)}{.447}\, hsGPA + \underset{(.0105)}{.0087}\, ACT$$
$$n = 141,\ R^2 = .219.$$

- Who is the treatment group and who is the control group?
- What is the t-stat for PC? How do you interpret 0.157?
- What if we drop ACT?
- What if we drop $hsGPA$?
- What if we use $NoPC$ instead of PC?

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).

Using dummy explanatory variables in equations for $log(y)$
(Percentage Interpretation)

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

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

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

Dummy indicating whether house is of colonial style

What does the coefficient of *colonial* mean?

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

As the dummy for colonial style changes from 0 to 1, the house price increases by 5.4 percent, holding all other factors fixed!

Example: Log Hourly Wage Equation

$$\widehat{\log(wage)} = \underset{(.099)}{.417} - \underset{(.036)}{.297}\, female + \underset{(.007)}{.080}\, educ + \underset{(.005)}{.029}\, exper$$
$$- \underset{(.00010)}{.00058}\, exper^2 + \underset{(.007)}{.032}\, tenure - \underset{(.00023)}{.00059}\, tenure^2$$
$$n = 526,\ R^2 = .441.$$

- What does the coefficient of $femal$ mean?

the coefficient on female implies that, for the same levels of $educ$, $exper$, and $tenure$, women earn about $100(.297) = 29.7\%$ less than men.

- What if we want to add a new dummy variable for $marriage$?

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)

Discussion:

- ✓ Interpret the dummy coefficients
- ✓ Difference between single and married women?
- ✓ Is this difference between single and married women statistically significant?

$$\widehat{\log(wage)} = \underset{(.106)}{.123} + \underset{(.056)}{.411}\, marrmale + \underset{(.058)}{.198}\, singmale + \underset{(.052)}{.088}\, singfem + \ldots,$$

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

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

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

$$n = 2{,}725, R^2 = .0422$$

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

Interactions involving dummy variables

- Allowing for different slopes

Interaction term

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

β_0 = intercept men

$\beta_0 + \delta_0$ = intercept women

β_1 = slope men

$\beta_1 + \delta_1$ = slope women

- Interesting hypotheses

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

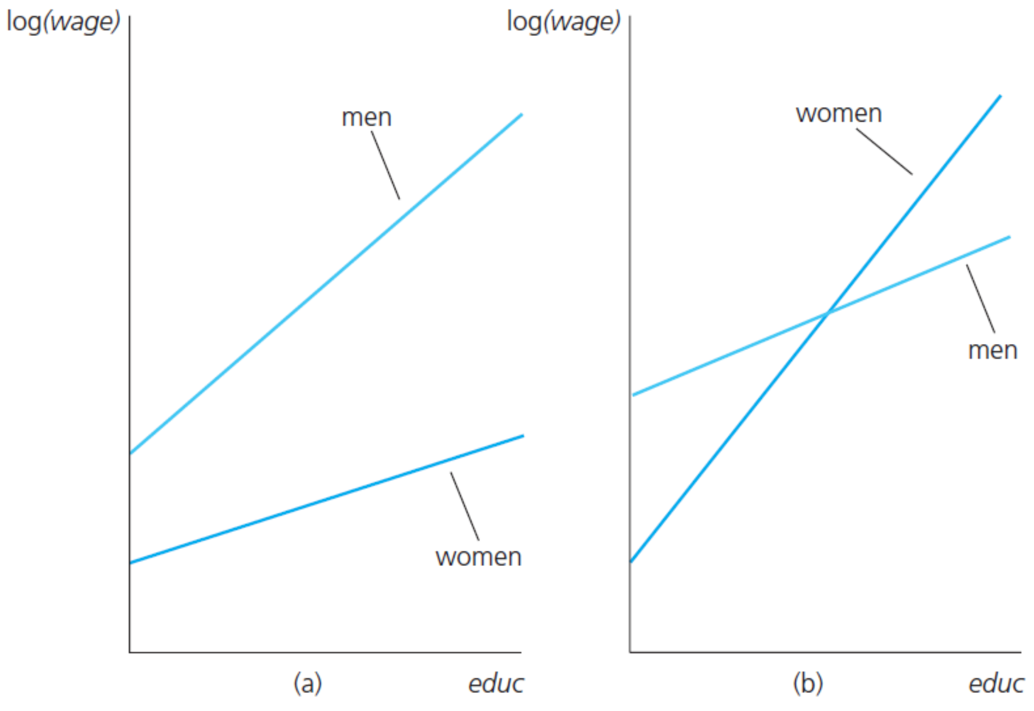
The return to education is the same for men and women

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

The whole wage equation is the same for men and women

Graphical illustration

FIGURE 7.2 Graphs of equation (7.16): (a) $\delta_0 < 0$, $\delta_1 < 0$; (b) $\delta_0 < 0$, $\delta_1 > 0$.



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

β_0 = intercept men

$\beta_0 + \delta_0$ = intercept women

β_1 = slope men

$\beta_1 + \delta_1$ = slope women

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

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

EXAMPLE 7.10 Log Hourly Wage Equation

We add quadratics in experience and tenure to (7.17):

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

Estimated return to education for **men (base group)**
What is the estimated return to education for women?

No evidence against hypothesis that the **return to education** is the same for men and women. Find the partial effect of education on log(wage)!

Does this mean that there is no significant evidence of lower pay for women at the same levels of educ, exper, and tenure? Hint: find the partial effect of being female on log(wage)!

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). HW9

Testing for differences in regression functions across groups

- ❑ Unrestricted model (contains full set of interactions)

College grade point average

High school rank percentile

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

$$+ \delta_2 \boxed{female} \cdot hsperc + \beta_3 tothrs + \delta_3 \boxed{female} \cdot tothrs + u$$

Total hours of colledge courses

Total hours spent in college courses

- ❑ Restricted model (same regression for both groups)

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

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

Testing for differences in regression functions across groups (cont'd)

- Estimation of the unrestricted model

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

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

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

Why we cannot conclude that cumgpa is about 0.353 less for women than for men?

Joint test with F-statistic

```
#case1: If both the intercept difference and the slope differences are zero.
linearHypothesis(MRM_dummy_UR, c("female=0", "female:sat=0" , "female:hsperc=0", "female:tothrs=0"))

Hypothesis:
female = 0
female:sat = 0
female:hsperc = 0
female:tothrs = 0

Model 1: restricted model
Model 2: cumgpa ~ female + sat + female:sat + hsperc + female:hsperc +
    tothrs + female:tothrs

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    362 85.515
2    358 78.355  4    7.1606 8.1791 2.545e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$cumgpa = \beta_0 + \delta_0 female + \beta_1 sat + \delta_1 female \cdot sat + \beta_2 hsperc + \delta_2 female \cdot hsperc + \beta_3 tothrs + \delta_3 female \cdot tothrs + u.$$

```
#case2: If the slope differences are zero.
linearHypothesis(MRM_dummy_UR, c("female:sat=0" , "female:hsperc=0", "female:tothrs=0"))

Hypothesis:
female:sat = 0
female:hsperc = 0
female:tothrs = 0

Model 1: restricted model
Model 2: cumgpa ~ female + sat + female:sat + hsperc + female:hsperc +
    tothrs + female:tothrs

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    361 79.362
2    358 78.355  3    1.0072 1.5339 0.2054
```

```r
stargazer(MRM_dummy_UR, MRM_dummy_R, type = "text", digits = 4)
```

```
=====================================================================
                                  Dependent variable:
                    -------------------------------------------------
                                        cumgpa
                              (1)                      (2)
---------------------------------------------------------------------
female                      -0.3535                 0.3101***
                            (0.4105)                (0.0586)

sat                        0.0011***                0.0012***
                            (0.0002)                (0.0002)

hsperc                     -0.0085***              -0.0084***
                            (0.0014)                (0.0012)

tothrs                     0.0023***                0.0025***
                            (0.0009)                (0.0007)

female:sat                  0.0008*
                            (0.0004)

female:hsperc               -0.0005
                            (0.0032)

female:tothrs               -0.0001
                            (0.0016)

Constant                   1.4808***                1.3285***
                            (0.2073)                (0.1798)

---------------------------------------------------------------------
Observations                  366                      366
R2                          0.4059                   0.3983
Adjusted R2                 0.3943                   0.3916
Residual Std. Error   0.4678 (df = 358)        0.4689 (df = 361)
F Statistic        34.9456*** (df = 7; 358) 59.7394*** (df = 4; 361)
=====================================================================
Note:                                     *p<0.1; **p<0.05; ***p<0.01
```

Using multiple categories vs. interaction terms

Model 1

$$\log(\text{wage}) \sim \text{mm} + \text{mf} + \text{sf} + \text{educ} + \text{exper} + \text{exper}^2 + \text{tenure} + \text{tenure}^2$$

```
reg_categories <- lm(lwage~I(married*(1-female))+ I(married*female) + I((1-married)*female) + educ + exper+I(exper^2) + tenure + I(tenure^2), wage1)
```

Model 2

$$\log(\text{wage}) \sim f + m + f * m + \text{educ} + \text{exper} + \text{exper}^2 + \text{tenure} + \text{tenure}^2$$

```
reg_interaction <- lm(lwage~ female + married + female:married + educ + exper+I(exper^2) + tenure + I(tenure^2), wage1)
```

Using multiple categories vs. interaction terms

- model 1 (using multiple categories) is better if you are interested in testing for wage differentials between any group and the base group

- model 2 (using interaction terms) allows us to easily test the null hypothesis that the gender differential does depend on marital status or not. Intercept significance vs slope significance.

	Dependent variable:	
	lwage	
	(1)	(2)
I(married * (1 - female))	0.213***	
	(0.055)	
I(married * female)	-0.198***	
	(0.058)	
I((1 - married) * female)	-0.110**	
	(0.056)	
female		-0.110**
		(0.056)
married		0.213***
		(0.055)
educ	0.079***	0.079***
	(0.007)	(0.007)
exper	0.027***	0.027***
	(0.005)	(0.005)
I(exper2)	-0.001***	-0.001***
	(0.0001)	(0.0001)
tenure	0.029***	0.029***
	(0.007)	(0.007)
I(tenure2)	-0.001**	-0.001**
	(0.0002)	(0.0002)
female:married		-0.301***
		(0.072)
Constant	0.321***	0.321***
	(0.100)	(0.100)
Observations	526	526
R2	0.461	0.461
Adjusted R2	0.453	0.453
Residual Std. Error (df = 517)	0.393	0.393
F Statistic (df = 8; 517)	55.246***	55.246***
Note:	*p<0.1; **p<0.05; ***p<0.01	