7 Multiple Regression Analysis with Qualitative Information

1. Which variables appear to reveal qualitative information? What are their features?

2. How do we understand regression with such variables and how do we study interactions of quantitative with such variables? What is the partial effect?

3. How binary dependent variables can be treated separately using our information on Binomial distributions?

4. Introducing natural experiments

7.1 Describing Qualitative Information

1. *Binary or zero-one or dummy variables*: describe experiments (or events) which have only two possible outcomes. Which variable is assigned as zero is a choice.

2. *Why only zero-one?* provides ease of interpretation.

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

7.2 Single Dummy Variable

1. Example: Consider the effect of education on wages.

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

Now we want to understand whether there are systematic differences wages across gender i.e. we want to study discrimination. We modify the model as

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

2. **Base group**: is the group which is denoted with 0. Here female = 1 when the observation refers to females and 0 if males, hence the base group is that of male. Your comparison would be with a male.

3. **Dummy variable trap**: we could instead code the variable male, when male = 1 when the observation refers to male, and 0 if females.

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 female + \beta_4 male + u$$

But including both in one model would include perfect collinearity and hence wrong.

4. **Discrimination on females in the labor market**: If we suspect this to be the case, we use this model with the null hypothesis that $H_0 : \beta_3 = 0$ i.e. there is no discrimination with alternative as $H_a : \beta_3 < 0$. Let us assume we estimated this regression:

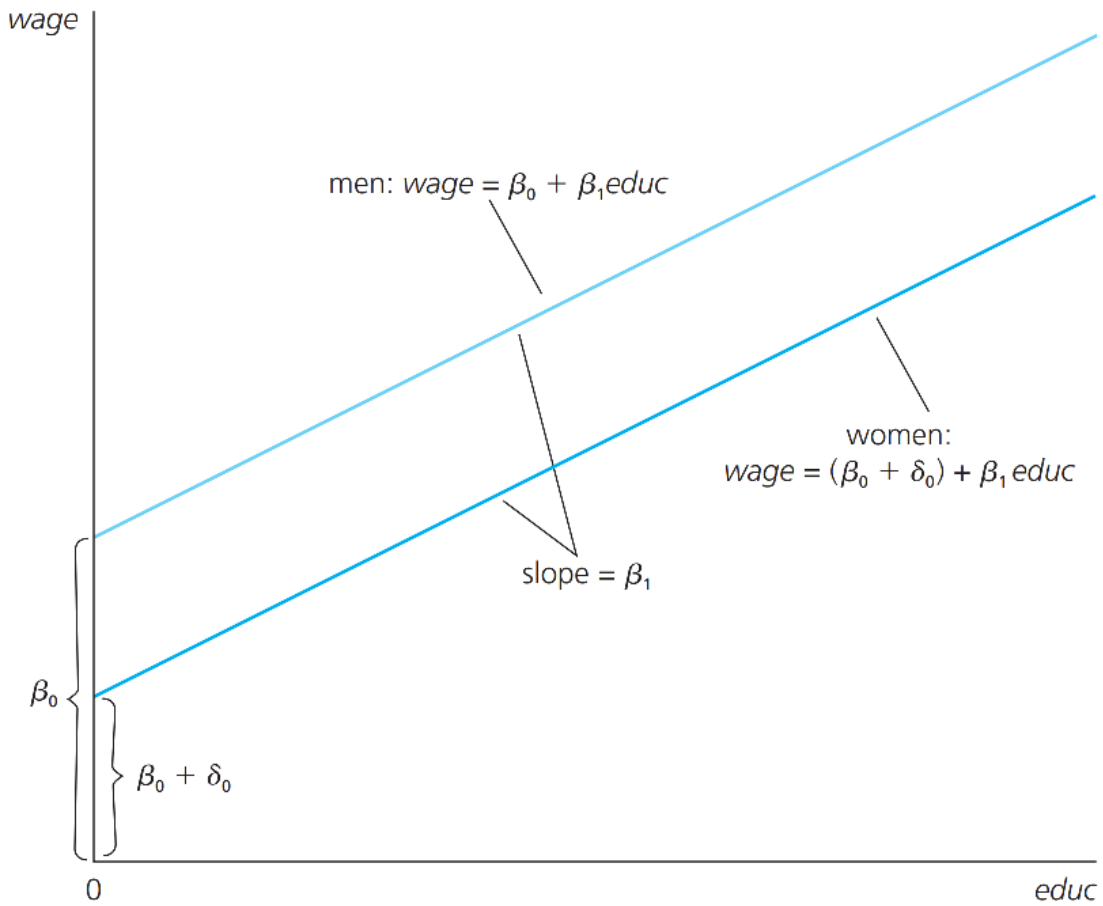
$$\hat{wage} = \hat{\beta}_0 + \hat{\beta}_1 educ + \hat{\beta}_2 exper + \hat{\beta}_3 female$$

5. **Interpretation**: magnitude of β_3 tells us the difference in hourly wage between females and males, given the same amount of education and experience. If the estimates are causal then:

$$\begin{aligned}\beta_3 =& E(wage|female = 1, educ, exper) \\ & - E(wage|female = 0, educ, exper)\end{aligned}$$

This is intercept shift between the fitted lines for male and female when the variables are coded as zero-one.

FIGURE 7.1 Graph of $wage = \beta_0 + \delta_0 female + \beta_1 educ$ for $\delta_0 < 0$.



6. Estimation results:

$$\begin{aligned} \hat{wage} = &- 1.57(0.72) + 0.572(0.049)educ \\ &+ 0.25(0.012)exper - 1.81(0.26)female \end{aligned}$$

(a) Females are paid 1.81$ less than men for every hour worked and none of these differences are explained by education or experience difference.

(b) **Nested model**: tells us that the average wage for men is 7.10$ in the sample while women earn only 4.59 without controlling for any other factors. This merely describes the sample and we do not claim causality.

$$\hat{wage} = 7.10(0.21) - 2.51(0.30)female$$

(c) **Inference**: for the t-stat on female to be valid, we assume that population variance in male is identical to females i.e. homoskedasticity.

7. Computer ownership and GPA: could be evaluated by examining a dummy where owns = 1 implies that student has a personal computer and owns = 0 implies student does not. *Here owns is a choice and not an attribute* and dummy variables could be used in this too.

8. **Program evaluation and Binary variables**: Programs are special cases of policy where there are likely to be two groups (a) control group: those who are not directly affected by policy and (b) treatment group i.e. those who are directly affected by policy.

   (a) Comes from natural sciences and experiments on rats etc.

   (b) Vaccine efficacy studies are a type of program evaluation where you choose people of similar characteristic and give vaccine to half (called the treated) and give the other half a placebo (called the control)

7.2.1 When Dependent Variable is $Log(y)$

1. Recall: when dependent variable is $log(y)$, coefficient on the dummy variable multiplied by 100 is interpreted as percentage difference in y (for the category dummied)

2. If the estimated $\hat{\beta}$ is large, then the we need a more accurate correction for computing the β associated with y when estimating a $log(y)$ model. This is again as previously undertaken in section 6.2

$$100\Big[exp(\hat{\beta}_1) - 1\Big]$$

7.3 Using Dummy Variable for Multiple Categories

1. Example on Marriage Premium: The question is whether there is discrimination of women and even more discrimination against married women. This is an example where there are 4 groups:

   (a) $Marrmale$, $marrfem$, $singfem$ with base group as single male

   (b) You could have taken any other base group too.

2. Estimates and interpretation:

$$\begin{aligned} log(\hat{wage}) =& 0.321^{**} + 0.213^{**} marrmale - 0.198^{**} marrfem - 0.110^{**} singfem \\ &+ 0.79^{**} educ + 0.027^{**} exper - 0.00054^{**} exper^2 \end{aligned}$$

   **: tells us that all variables are significant at 5% level. Married men earn about 21.3% more than single male (base group)

3. **Discrimination against married women**: is of the order of 19.8% in comparison to single males of the same education and experience.

4. Difference between groups with correct standard errors:

   - We can approximate the difference between single and married women as (-0.110 - (-0.198) = 0.088) i.e. 9% difference. But we dont know the correct SD on this for testing. We can re-estimate the model with married females as base to directly obtain SD.

7.3.1 Incorporating Ordinal Information by Using Dummy Variables

1. *Example: Scale of Physical Attractiveness*

   Consider five categories i.e. homely, quite plain, average, good looking, and strikingly beautiful (or handsome). **A survey was conducted asking independent experts in fashion about employees physical appearance** (withholding their name, designation etc.)

   Since there were very few extreme observations i.e. homely and strikingly handsome, **the categories were renamed** as : average, below and above.

2. Estimated equation is:

$$\text{Men :} \qquad \hat{log(wage)} = \hat{\beta}_0 - 0.164^{**}belavg + 0.016abvavg + ...$$

$$\text{Women :} \qquad \hat{log(wage)} = \hat{\beta}_0 - 0.124^{*}belavg + 0.035abvavg + ...$$

3. **Interpretation**: a man who appears not good looking obtains a 16% less wage and a similar wage for a below average women is 12%. This is discrimination on physical appearance against men.

   Hamermesh, D. S., & Biddle, J. E. (1993). Beauty and the labor market (No. w4518). National Bureau of Economic Research.

4. Too many ordinal variables: could be reduced into "arbitrary" categories and the used with similar interpretation, albeit with caution.

   ▷ Law school example 7.8