QUALITATIVE EXPLANATORY VARIABLES in REGRESSION ANALYSIS

Hüseyin Taştan[1]

[1]Yıldız Technical University
Department of Economics

Econometrics I

Qualitative Information in Regression Analysis

- Two kinds of variables: quantitative vs. qualitative
- So far we only used quantitative information in our regression models, e.g., wages, experience, house prices, number of rooms, GPA, attendance rate, etc.
- In practice we would like to include qualitative variables in the regression.
- For example: gender, ethnicity, religion of an individual, region or location of an individual or city, industry of a firm (manufacturing, retail, finance,...) etc.
- This kind of categorical variables can be represented by binary or dummy variables.

Qualitative Variables: Lecture Plan

- Describing Qualitative Information
- A Single Dummy Independent Variable
- Dummy Variables for Multiple Categories
- Interactions Involving Dummy Variables
- Binary Dependent Variable (Linear Probability Model)

Qualitative Information

- In most cases qualitative factors come in the form of binary information: female/male, domestic/foreign, north/south, manufacturing/nonmanufacturing, countries with or without capital punishment laws, etc.
- Dummy variables: also called binary (0/1) variable.
- Any kind of categorical information can easily represented by dummy variables.
- It does not matter which category is assigned the value 0 or 1. But we need to know the assignment to interpret the results.
- For example: gender dummy in the wage equation: female=1, male=0.
- Marital status: married=1, single=0.
- Location of the country: northern hemisphere=1, southern hemisphere=0

Example Data Set: wage1.gdt

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

Single Dummy Independent Variable

- How to include binary information into regression model?
- Let one of the x variables be a dummy variable:

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

- For female workers $female = 1$ for male worker $female = 0$.
- How to interpret δ_0: the difference in hourly wage between females and males, given the same amount of education (and the same error term u).
- Is there discrimination against women in the labor market?
- If $\delta_0 < 0$ then we will be able to say that given the same level of education female workers earn less than male workers on average.
- This can easily be tested using t-statistic.

Single Dummy Independent Variable

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

- Conditional expectation of wage for women:

$$\mathrm{E}(wage|female = 1, educ) = \beta_0 + \delta_0 + \beta_1\ educ$$

- For men:

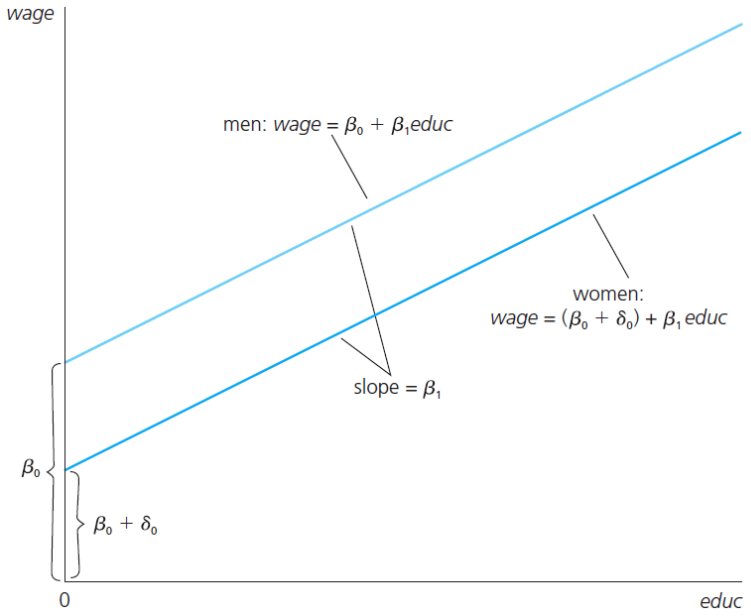
$$\mathrm{E}(wage|female = 0, educ) = \beta_0 + \beta_1\ educ$$

- Taking the difference:

$$\mathrm{E}(wage|female = 1, educ) - \mathrm{E}(wage|female = 0, educ)$$

$$= \beta_0 + \delta_0 + \beta_1\ educ - (\beta_0 + \beta_1\ educ) = \boldsymbol{\delta_0}$$

Wage Equation for $\delta_0 < 0$



Single Dummy Independent Variable

- In the wage equation β_0 is the intercept term for male workers (let female=0).
- The intercept term for the female workers is $\beta_0 + \delta_0$.
- A single dummy variable can differentiate between two categories. We do not need to include a separate dummy variable for males.
- In general: the number of dummy variables = the number of categories minus 1
- In the wage equation we have just two groups. Using two dummy variables would introduce perfect collinearity because $female + male = 1$.
- This is called **dummy variable trap**.
- Dummy=0 is called the **base group** or benchmark group. This is the group against which comparisons are made. In the formulation above the base group is male workers.
- The coefficient on *female* (δ_0) gives the difference in intercepts between females and males.

Single Dummy Independent Variable

- Male workers as the base group: $female = 1$ for female workers

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

- Female workers as the base group: $male = 1$ for male workers

$$wage = \alpha_0 + \gamma_0\ male\ + \beta_1\ educ + u$$

- Intercept for female workers: $\alpha_0 = \beta_0 + \delta_0$
- Intercept for male workers: $\alpha_0 + \gamma_0 = \beta_0$
- We need to know which group is the base group.

Single Dummy Independent Variable

- Another alternative is to write the model without the intercept term and including dummy variables for each group:

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

- No dummy variable trap as there is no intercept.
- Notice that coefficients on dummies give us intercepts for each group.
- We do not prefer this specification because it is not clear how to calculate R^2. It may even be negative.
- Also, testing for a difference in intercepts is more difficult.

Adding Quantitative Variables

- Adding quantitative variables does not change the interpretation of dummy variables. Consider the following model with male workers as the base group:

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

- δ_0: Intercept difference between female and male workers at the same level of education, experience and tenure.
- Testing for discrimination: $H_0 : \delta_0 = 0$ vs $H_1 : \delta_0 < 0$
- If we reject H_0 in favor of the alternative there is evidence of discrimination against women in the labor market.
- Can easily be tested using t statistic.

Example: Wage Equation

$$\widehat{\text{wage}} = \underset{(0.725)}{-1.57} - \underset{(0.265)}{1.81}\,\text{female} + \underset{(0.049)}{0.572}\,\text{educ} + \underset{(0.0116)}{0.025}\,\text{exper} + \underset{(0.021)}{0.141}\,\text{tenure}$$

$$n = 526 \quad R^2 = 0.364 \quad F(4, 521) = 74.398 \quad \hat{\sigma} = 2.9576$$

- On average, women earn \$1.81 less than men, ceteris paribus. More specifically, if we take a woman and a man with the same levels of education, experience and tenure, the woman earns, on average, \$1.81 less per hour than the man.
- -1.57: this is the intercept for male workers. Not meaningful as there is no one in the sample with zero values of education, experience and tenure.

Logarithmic Dependent Variable

$$\widehat{\log(\text{wage})} = \underset{(0.099)}{0.417} - \underset{(0.036)}{0.297}\,\text{female} + \underset{(0.007)}{0.080}\,\text{educ} + \underset{(0.005)}{0.029}\,\text{exper} - \underset{(0.0001)}{0.00058}\,\text{expersq} + \underset{(0.007)}{0.032}\,\text{tenure} - \underset{(0.00023)}{0.00059}\,\text{tenursq}$$

$$n = 526 \quad R^2 = 0.441 \quad F(6, 519) = 68.177 \quad \hat{\sigma} = 0.39978$$

(standard errors in parentheses)

- Interpretation of the coefficient on female: women earn about $100 \times 0.297 = 29.7\%$ less than men for the same levels of educ, exper, and tenure.
- We can compute a more accurate approximation for the proportionate difference in wages between men and women holding all other factors fixed.

$$\frac{\widehat{wage}_F - \widehat{wage}_M}{\widehat{wage}_M} = \exp(-0.297) - 1 \approx -0.257$$

- Women earn approximately 25.7% less than comparable men.

Dummy Variables: No quantitative variable in the regression

Suppose that we exclude all quantitative variables from the model:

$$\widehat{\text{wage}} = \underset{(0.21)}{7.1} - \underset{(0.303)}{2.51}\,\text{female}$$

$$n = 526 \quad \bar{R}^2 = 0.1140$$

- The intercept is simply the average wage for men in the sample (7.1\$).
- Coefficient estimate on female: the difference in the average wage between women and men (\$2.51).
- The average wage for women in the sample is: $7.1 - 2.51 = \$4.59$
- If we calculate the sample averages for each group we will get the same results. Notice that we did not control for any explanatory variables in this case.

Dummy Variables: No quantitative variables

$$\widehat{\text{wage}} = \underset{(0.21)}{7.1} - \underset{(0.303)}{2.51}\,\text{female}$$

$$n = 526 \quad \bar{R}^2 = 0.1140$$

- The model above can be used to compute the simple comparison-of-means test between the two groups (in our example between two genders)
- This is just a simple t-test on the dummy variable: $t = -2.51/0.303 = -8.28$ which is larger than the critical value at any reasonable significance level. Thus, the evidence suggests that the means across groups are not the same.
- The comparison-of-means t-test is valid under the assumption of homoscedasticity. If the variances are different across groups then we should use appropriate correction.
- We again note that the model that includes additional factors (education, experience, tenure, etc.) in the model is more appropriate to estimate the ceteris paribus gender wage gap.

More Than One Dummy Variables

Let us define two dummy variables: $female = 1$ if the worker is female; $married = 1$ if the worker is married

$$\widehat{\text{wage}} = \underset{(0.296)}{6.18} - \underset{(0.302)}{2.29}\ \text{female} + \underset{(0.310)}{1.34}\ \text{married}$$

$$n = 526 \quad \bar{R}^2 = 0.1429$$

- Coefficient on female is just the intercept difference between female workers and male workers holding marital status fixed: -2.29.
- Similarly, the coefficient on married is the intercept difference between single and married workers (regardless of gender).
- Note that in this equation the marriage differential is assumed to be the same across genders.
- We can add an interaction variable $female \times married$

More Than One Dummy Variables

$$\widehat{\text{wage}} = \underset{(0.361)}{5.168} - \underset{(0.474)}{0.556}\,\text{female} + \underset{(0.436)}{2.815}\,\text{married} - \underset{(0.608)}{2.861}\,\text{female} \times \text{married}$$

$$n = 526 \quad \bar{R}^2 = 0.18$$

- According to the results above, married men earn \$2.815 more than single men on average.
- On the other hand, married women earn 0.60 USD less than single man (note that $-0.556 + 2.815 - 2.861 = -0.601$) on average. The sample average of the married & female group is \$4.567 ($= 5.168 - 0.601$)
- We need to control for relevant quantitative variables (education, experience, tenure, etc.) so that we can use the ceteris paribus notion.

Wage Equation

$$\widehat{\text{lwage}} = \underset{(0.098)}{0.42} - \underset{(0.036)}{0.29}\ \text{female} + \underset{(0.040)}{0.05}\ \text{married} + \underset{(0.007)}{0.08}\ \text{educ}$$
$$+ \underset{(0.005)}{0.03}\ \text{exper} - \underset{(0.0001)}{0.0005}\,\text{expersq} + \underset{(0.007)}{0.03}\ \text{tenure} - \underset{(0.0002)}{0.0006}\,\text{tenursq}$$

$$n = 526 \quad \bar{R}^2 = 0.4351$$

- After controlling for the other factors is there still difference in average wages between single male workers and married male workers?
- Coefficient on married: 0.05. Associated t statistic: $0.05/0.04 = 1.25$. Fail to reject H_0.
- But in the model above, we assumed that the marriage premium is the same for men and women. To relax this, we can just add an interaction dummy (just like we did in the previous model) or create a set of dummy variables corresponding to 2×2 classification based on female and married dummies.

Dummy Variables for Multiple Categories

- Using $female$ and $married$ we can separate workers into 4 groups and define dummy variables for these groups as follows:

$$marrmale = married \times (1 - female)$$

$$marrfem = married \times female$$

$$singfem = (1 - married) \times female$$

$$singmale = (1 - married) \times (1 - female)$$

- $marrmale$ is the dummy for the married male workers, $marrfem$ married female workers, $singfem$: single female workers and $singmale$ is the single male workers.
- Need to choose one of these as the base group so that we include $4 - 1 = 3$ dummies in the model.
- Suppose that the base group is $singmale$.

Dummy Variables for Multiple Categories

$$\widehat{\text{lwage}} = \underset{(0.101)}{0.32} + \underset{(0.055)}{0.21}\,\text{marrmale} - \underset{(0.058)}{0.198}\,\text{marrfem} - \underset{(0.055)}{0.11}\,\text{singfem} + \underset{(0.006)}{0.079}\,\text{educ} + \underset{(0.005)}{0.027}\,\text{exper} - \underset{(0.0001)}{0.0005}\,\text{expersq} + \underset{(0.006)}{0.029}\,\text{tenure} - \underset{(0.0002)}{0.0005}\,\text{tenursq}$$

$$n = 526 \quad R^2 = 0.461 \quad F(8, 517) = 55.246 \quad \hat{\sigma} = 0.39329$$

- Coefficient on $marrmale$: 0.21: Married men are estimated to earn about 21% more than single men (proportionate difference relative to the base group which is single male), holding all other factors fixed.
- A married women earns 19.8% less than a single man with the same levels of the other variables.

Dummy Interactions

Recast the model using the interaction of female and married

$$\widehat{\text{lwage}} = \underset{(0.100)}{0.321} - \underset{(0.056)}{0.110}\,\text{female} + \underset{(0.055)}{0.213}\,\text{married} - \underset{(0.072)}{0.301}\,\text{female} \times \text{married} + \underset{(0.0067)}{0.079}\,\text{educ} + \underset{(0.0052)}{0.027}\,\text{exper} - \underset{(0.00011)}{0.000535}\,\text{expersq} + \underset{(0.0068)}{0.0291}\,\text{tenure} - \underset{(0.00023)}{0.00053}\,\text{tenursq}$$

$$n = 526 \quad R^2 = 0.461 \quad F(8, 517) = 55.246 \quad \hat{\sigma} = 0.39329$$

- When we set female=0, married = 0 we obtain the group single male workers (interaction is automatically zero in this case). This is the base group (and the intercept is 0.321).
- The coefficient on married gives the intercept difference between married and single men (which is about 21%, the same as before).

Dummy Interactions

$$\widehat{\text{lwage}} = \underset{(0.100)}{0.321} - \underset{(0.056)}{0.110}\,\text{female} + \underset{(0.055)}{0.213}\,\text{married} - \underset{(0.072)}{0.301}\,\text{female} \times \text{married} + \underset{(0.0067)}{0.079}\,\text{educ} + \underset{(0.0052)}{0.027}\,\text{exper} - \underset{(0.00011)}{0.000535}\,\text{expersq} + \underset{(0.0068)}{0.0291}\,\text{tenure} - \underset{(0.00023)}{0.00053}\,\text{tenursq}$$

$$n = 526 \quad R^2 = 0.461 \quad F(8, 517) = 55.246 \quad \hat{\sigma} = 0.39329$$

- Intercepts for other groups can easily be found, for example, the intercept for married women is $0.32 - 0.11 + 0.213 - 0.301 = 0.122$. In other words, married women earn about 19.9% less than comparable men.
- This specification is essentially the same as the previous one. The explanatory power is the same (look at the Rsq and and F stats) But this specification can be used to test if gender differential depends on marital status and vice versa.

Allowing for Different Slopes using Interaction Terms

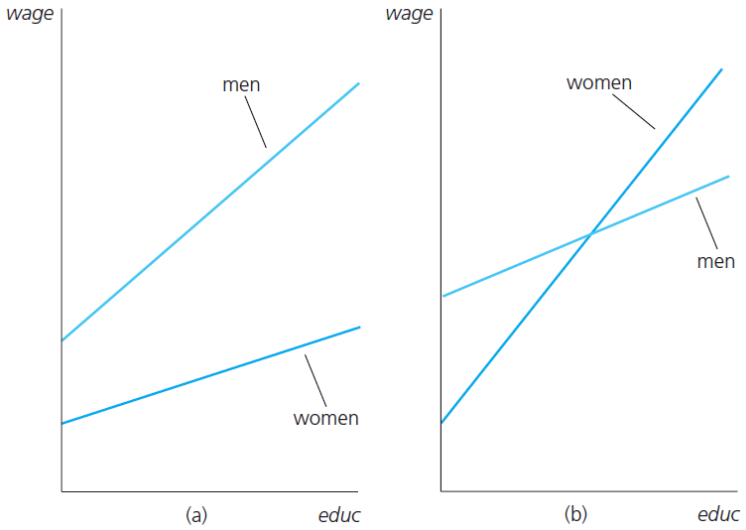
- So far we assumed that slope coefficients on the quantitative variables are constant but intercepts are different. In some cases we want to allow for different slopes as well as different intercepts.
- For example, suppose that we want to test whether the return to education is the same for men and women.
- To estimate different slopes it suffices to include an interaction term involving $female$ and $educ$: $female \times educ$.

Allowing for Different Slopes: Wage Equation

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

- Plugging in $female = 0$ we see that β_0 is the intercept for male workers.
- β_1 is the slope on education for males.
- Plugging in $female = 1$, δ_0 is the difference between intercepts for female and male workers. Thus, the intercept term for females is $\beta_0 + \delta_0$
- δ_1 is measures the difference in the return to education between women and men. Slope on education for female: $\beta_1 + \delta_1$
- If $\delta_1 > 0$ then we can say that the return to education for women is larger than the return to education for men.

Allowing for Different Slopes: Left: $\delta_0 < 0, \delta_1 < 0$; Right: $\delta_0 < 0, \delta_1 > 0$



Difference in Slopes for the Wage Equation

- Graph (a): the intercept for women is below that for men, and the slope of the line is smaller for women than for men.
- This means that women earn less than men at all levels of education and the gap increases as $educ$ gets larger.
- Graph (b): the intercept for women is below that for men, but the slope on education is larger for women.
- This means that women earn less than men at low levels of education, but the gap narrows as education increases.
- At some point, a women earns more than a man given the same level of education.

Interaction between Gender and Education

The model can be formulated as follows:

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

- We just added $female \times educ$ interaction term along with $female$ and $educ$.
- Interaction variable will be 0 for men, and $educ$ women.
- $H_0 : \delta_1 = 0,\ H_1 : \delta_1 \neq 0$. This says "The return to another year of education is the same for men and women"
- $H_0 : \delta_0 = 0, \delta_1 = 0$: "Average wages are identical for men and women who have the same levels of education". Carry out an F test.

Interaction between Gender and Education

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

Interaction between Gender and Education

- Estimated return to education for men is 8.2%.
- For women, return to education is $0.082 - 0.0056 = 0.0764$, or about 7.6%. The difference, given by the interaction coefficient, is -0.56%
- This is not economically large and statistically insignificant: t statistic is $-0.0056/0.0131 = -0.43$.
- Coefficient on $female$ measures the wage difference between men and women when $educ = 0$.
- Note that there is no one with 0 years of education in the sample. Also, due to high collinearity between $female$ and $female \cdot educ$ its standard error is high and t ratio is small (-1.35).

Interactions Involving Dummy Variables

- Instead of omitting $female$ we will estimate its coefficient by redefining the interaction term.
- Instead of interacting $female$ with $educ$ we will interact it with the deviation from the mean education level. Average education level in the sample is 12.5 years
- Our new interaction term is: $female \times (educ - 12.5)$.
- In this regression, the coefficient on female will measure the average wage difference between women and men at the mean education level, $educ = 12.5$.

Example: Wage Equation, R output

```
> data(wage1, package='wooldridge')
> # interaction: female*(educ-12.5)
> wage1$educdev <- (wage1$educ-12.5)
> res2 <- lm(lwage ~ female*educdev + exper + I(exper^2) + tenure + I(tenure^2), data=wage1)
> summ(res2, digits = 5)

MODEL INFO:
Observations: 526
Dependent Variable: lwage
MODEL FIT:
F(7,518) = 58.37084, p = 0.00000
Rsq = 0.44096
Adj. Rsq = 0.43341
-----------------------------------------------------------------
                       Est.      S.E.     t val.          p
-------------------- ---------- --------- ---------- ---------
(Intercept)            1.41842   0.04405   32.20343   0.00000
female                -0.29635   0.03584   -8.26952   0.00000
educdev                0.08237   0.00847    9.72492   0.00000
exper                  0.02934   0.00498    5.88597   0.00000
I(exper^2)            -0.00058   0.00011   -5.39777   0.00000
tenure                 0.03190   0.00686    4.64696   0.00000
I(tenure^2)           -0.00059   0.00024   -2.50890   0.01242
female:educdev        -0.00556   0.01306   -0.42601   0.67028
-----------------------------------------------------------------
```

$$\begin{aligned}lwage = 1.41842 - 0.29635(female) + 0.08237(educ - 12.5) + 0.02934(exper) - \\ 0.00058(exper^2) + 0.0319(tenure) - 0.00059(tenure^2) - 0.00556 female \times (educ - 12.5) + resid\end{aligned}$$