QUALITATIVE INFORMATION in REGRESSION ANALYSIS

Hüseyin Taştan[1]

[1]Yıldız Technical University
Department of Economics

These presentation notes are based on
Introductory Econometrics: A Modern Approach (2nd ed.)
by J. Wooldridge.

1 Aralık 2012

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:

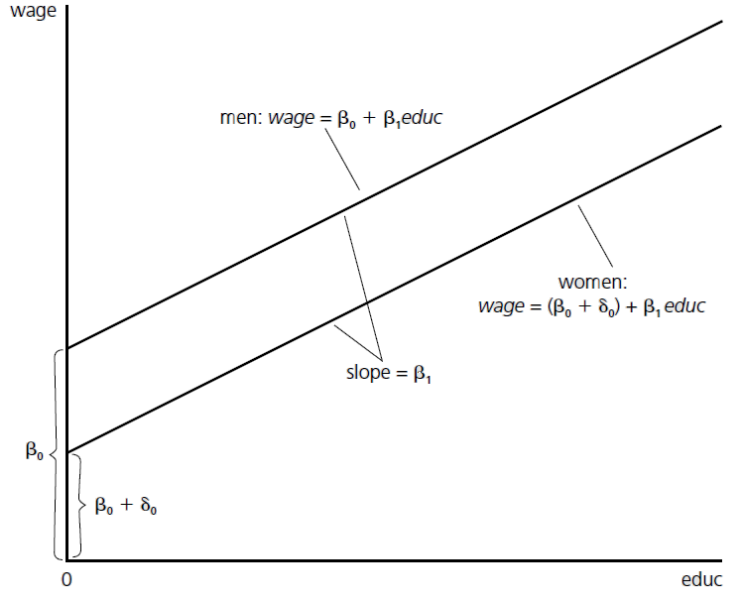
$$\mathrm{E}(wage|female = 0, educ) = \beta_0 + \beta_1\ educ$$

- 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.

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.

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

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

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

- Base group "single male workers" $(female = 0, married = 0)$. Intercept estimate for the base group is $\$6.18$.
- Coefficient on female is just the average wage difference between female workers and single male workers: $\$2.51$.
- What is the average wage for single female workers? $(female = 1, married = 0)$: $6.18 - 2.29 = \$3.89$

More Than One Dummy Variables

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

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

- Similarly, average wage for married female workers $(female = 1, married = 0)$: $6.18 - 2.29 + 1.34 = \$5.23$
- Average wage difference between married males and married females: $(6.18 + 1.34) - (6.18 - 2.29 + 1.34) = \$2.29$. Men earn more than women on average.
- 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.

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

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

$$n = 526 \quad \bar{R}^2 = 0.4525 \quad F(8, 517) = 55.246 \quad \hat{\sigma} = 0.39329$$

- 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.

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

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

$$n = 526,\ R^2 = .441.$$

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, STATA Output

```
. gen femeduc1=female*(educ-12.5)

. reg  lwage female educ femeduc1 exper expersq tenure tenursq

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  7,   518) =   58.37
       Model |  65.4081534     7  9.34402192           Prob > F      =  0.0000
    Residual |  82.921598    518  .160080305           R-squared     =  0.4410
-------------+------------------------------           Adj R-squared =  0.4334
       Total |  148.329751   525   .28253286           Root MSE      =   .4001

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -.296345   .0358358    -8.27   0.000    -.3667465   -.2259436
        educ |   .0823692   .0084699     9.72   0.000     .0657296    .0990088
    femeduc1 |  -.0055645   .0130618    -0.43   0.670    -.0312252    .0200962
       exper |   .0293366   .0049842     5.89   0.000      .019545    .0391283
     expersq |  -.0005804   .0001075    -5.40   0.000    -.0007916   -.0003691
      tenure |   .0318967    .006864     4.65   0.000      .018412    .0453814
    tenursq  |    -.00059   .0002352    -2.51   0.012    -.001052    -.000128
       _cons |   .388806   .1186871     3.28   0.001     .1556388    .6219732
------------------------------------------------------------------------------

. test  female femeduc1

 ( 1)  female = 0
 ( 2)  femeduc1 = 0

       F(  2,   518) =   34.33
            Prob > F =    0.0000
```

Binary Dependent Variable: Linear Probability Model

- So far, in all of the models we examined the dependent variable y has been a quantitative variable, e.g., wages, GPA score, prices, etc.
- Can we explain a qualitative (ie binary or dummy) variable using multiple regression?
- Binary dependent variable $y = 1$ or $y = 0$; eg it may indicate whether an adult has a high school education, whether a household owns a house, whether an adult is married, owns a car, etc.
- The case where $y = 1$ is called success whereas $y = 0$ is called failure.
- What happens if we regress a $0/1$ variable on a set of independent variables? How can we interpret regression coefficients?

Binary Dependent Variable: Linear Probability Model

- Under the standard assumptions the conditional expectation of the dependent variable can be written as follows:

$$\mathsf{E}(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

- Since y takes only values of 0 or 1 this conditional expectation can be written as follows:

$$\begin{aligned} \mathsf{E}(y|x) &= P(y = 1|x) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \end{aligned}$$

- The probability of success is given by $p(x) = P(y = 1|x)$. The expression above states that the success probability is a linear function of x variables.
- By definition the probability of failure is $P(y = 0|x) = 1 - P(y = 1|x)$

Binary Dependent Variable: Linear Probability Model

- Linear Probability Model (LPM):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

- x variables can be qualitative or quantitative.
- Slope coefficients are now interpreted as the change in the probability of success:

$$\Delta P(y = 1|x) = \beta_j \Delta x_j$$

- OLS sample regression function is given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

- $\hat{y}$ is the predicted probability of success.

Example: Women's Labor Force Participation, mroz.gdt

- y ($inlf$ - in the labor force) equals 1 if a married woman reported working for a wage outside the home in 1975, and 0 otherwise..
- Definitions of explanatory variables
- $nwifeinc$: husband's earnings (in \$1000),
- $kidslt6$: number of children less than 6 years old,
- $kidsge6$: number of children between 6-18 years of age,
- $educ, exper, age$
- Model

$$\widehat{inlf} = \hat{\beta}_0 + \hat{\beta}_1 nwifeinc + \hat{\beta}_2 educ + \ldots + \hat{\beta}_7 kidsge6$$

Women's Labor Force Participation, mroz.gdt

Model 1: OLS, using observations 1–753
Dependent variable: inlf

	Coefficient	Std. Error	t-ratio	p-value
const	0.585519	0.154178	3.7977	0.0002
nwifeinc	−0.00340517	0.00144849	−2.3508	0.0190
educ	0.0379953	0.00737602	5.1512	0.0000
exper	0.0394924	0.00567267	6.9619	0.0000
expersq	−0.000596312	0.000184791	−3.2270	0.0013
age	−0.0160908	0.00248468	−6.4760	0.0000
kidslt6	−0.261810	0.0335058	−7.8139	0.0000
kidsge6	0.0130122	0.0131960	0.9861	0.3244

Mean dependent var	0.568393	S.D. dependent var	0.495630
Sum squared resid	135.9197	S.E. of regression	0.427133
R^2	0.264216	Adjusted R^2	0.257303
$F(7, 745)$	38.21795	P-value(F)	6.90e–46

Women's Labor Force Participation, mroz.gdt

- All variables are individually statistically significant except $kidsge6$. All coefficients have expected signs using standard economic theory and intuition.
- Interpretation of coefficient estimates: For example, the coefficient estimate on $educ$, 0.038, implies that, ceteris paribus, an additional year of education increases predicted probability of labor force participation by 0.038.
- The coefficient estimate on $nwifeinc$: if husband's income increases by 10 units (ie, $\$10000$) the probability of labor force participation falls by 0.034.
- $exper$ has a quadratic relationship with $inlf$: the effect of past experience on the probability of labor force participation is diminishing.

Women's Labor Force Participation, mroz.gdt

- The number of young children has a big impact on labor force participation. The coefficient estimate on $kidslt6$ is -0.262.
- Ceteris paribus, having one additional child less than six years old reduces the probability of participation by -0.262.
- In the sample, about 20% of the women have at least one child.

Shortcomings of LPM

- Predicted probability of success is given by $\hat{y}$ and it can have values outside the range 0-1. Obviously, this contradicts the rules of probability.
- In the example out of 753 observations, 16 have $\widehat{inlf} < 0$ and 17 have $\widehat{inlf} > 1$.
- If these are relatively few, they can be interpreted as 0 and 1, respectively.
- Nevertheless, the major shortcoming of LPM is not implausible probability predictions. The major problem is that a probability cannot be linearly related to the independent variables for all their possible values.

Shortcomings of LPM

- In the example, the model predicts that the effect of going from zero children to one young child reduces the probability of working by 0.262.
- This is also the predicted drop if the woman goes from having one child to 2 or 2 to 3, etc.
- It seems more realistic that the first small child would reduce the probability by a large amount, but subsequent children would have a smaller marginal effect.
- Thus, the relationship may be nonlinear.

Shortcomings of LPM

- Despite these shortcomings LPM is useful and often applied in economics.
- It usually works well for values of the independent variables that are near the averages in the sample.
- In the previous example, 96% of the women have either no children or one child under 6. Thus, the coefficient estimate on $kidslt6$ (-0.262) practically measures the impact of the first children on the probability of labor force participation.
- Therefore, we should not use this estimate for changes from 3 to 4 or 4 to 5, etc.

Shortcomings of LPM

- LPM is heteroscedastic: The MLR.5: Constant error variance assumption is not satisfied.
- Recall that y is a binary variable following a Bernoulli distribution. Thus, the variance for a Bernoulli distribution is given by:

$$\text{Var}(u|x) = \text{Var}(y|x) = p(x) \cdot [1 - p(x)]$$

- Since $p(x)$ is a linear combination of x variables, $\text{Var}(u|x)$ is not constant.
- We learned that in this case OLS is unbiased and consistent but inefficient. The Gauss-Markov Theorem fails. Standard errors and the usual inference procedures are not valid.
- It is possible to find more efficient estimators than OLS.