ECONOMETRIC MODEL WITH QUALITATIVE VARIABLES

How to quantify qualitative variables to quantitative variables ?

Why do we need to do this ?
Econometric model needs quantitative variables to estimate its parameters

What are the differences among these variables:
Dummy? Indicator? Binary? Dichotomy? Categorical

ECONOMETRIC MODEL WITH DUMMY VARIABLES

Specifically:

What if the variables are not quantitative variables, like:

I. Male-female; Urban-rural; Yes-No; foreign-domestic
II. Level of education: SD, SLTP, SLTA, D3, S1, S2, S3
    Choice if investment: stock, certificate of BI, gold, etc.

Other Usages:

How to model Unstable Regression?

- Jumping Regression
- Shifting Regression

Technically speaking, do we have problems with our model if:

- Independent variable (s) is (are) a dummy (ies)
- Dependent variables is a dummy

Illustration:

We would like to analyze whether there are differences between graduate and undergraduate students in weekly entertainment spending.

Y: weekly spending for entertainment per student

PS: graduate or undergraduate

PS = 1 ; graduate student

PS = 0 ; undergraduate student

Model: $Y = \alpha + \beta PS + u$

From the model, an average spending:

- Graduate student: $E(Y \mid PS = 1) = \alpha + \beta$
- Undergraduate student: $E(Y \mid PS = 0) = \alpha$

For example, by using data from a survey, the estimated model is the following:

$$Y = 9{,}4 + 16\,PS$$
$$t \quad (53{,}22) \; (6{,}245)$$
$$R^2 = 96{,}54\%$$

The model indicates that $\alpha \neq 0$ dan $\beta \neq 0$ (statistically signifiant)

Interpretation:

average spending for graduate students: 9,4 + 16 = 25,4,

average spending for under graduate students: 9,4

(There is a difference between spending of the two groups)

The next question is whether graduate students more able or more consumptive in entertainment spending than undergraduate students

Professor's salary = f (experience, sex)

Do we have a discrimination in salary policy against female professors?

Y = yearly salary of a professor
X = years of teaching
G = 1 ; male professor
0 ; female professor

A model that can relate X and G to Y:

$$Y = \alpha_1 + \alpha_2 G + \beta X + u$$

From the model, it can be seen that:

- Average salary of female professor = $\alpha_1 + \beta X$
- Average salary of male professor = $\alpha_1 + \alpha_2 + \beta X$

Any discrimination against female professors?

Supposed, based on a survey, the estimated model:

$$Y = 19.21 + 0.373\ G + 1.453\ X$$

t: (11.33) (1.141) (37.997)

$$R^2 = 89.75\%$$

How about if we define dummy differently?

S = 1; female professor

= 0; male professor

Since we define dummy variable differently, will we have different result substantively?

Model with new definition:

$$Y = \alpha_1 + \alpha_2 S + \beta X + u$$

Remark

In defining dummy variable, which category is representing by one or zero does not matter as long as the estimated model is interpreted consistently.

What happened if we define dummy variable as follows:

D_2 = 1; male professor
0; female professor

D_3 = 1; female professor
0; male professor

The model with this definition:

$$Y = \alpha_1 + \alpha_2 D_2 + \alpha_3 D_3 + \beta X + u$$

When we estimate this model with OLS, what will happened ?

Qualitative Variables with more than two categories

Levels of Education: SD, SLTP, SLTA, D3, S1, S2, S3
Choices of Investments: Stock, Saving Deposits, Property, Gold

Can we represent these types of variables with dummy variables? How?

Supposed we have 3 categories of Education Levels:

(i) Graduate from Secondary School or lower,
(ii). Graduate from High School,
(iii). Graduate from University

Can we represent these types of variables with a Variable that has different values like: 1, 2, and 3 based on the number of categories?

Should we define differently?

Try define as follows:

D_2 = 1 ; if the highest level of education is high school
0 ; others

D_3 = 1 ; if the highest level of education is university
0 ; others

Do we need to define the other category explicitly?

Life Insurance Consumption = f (income, education)

See the following model:

$Y = \alpha_1 + \alpha_2 D_2 + \alpha_3 D_3 + \beta X + u$

Y = life insurance expenses per year
X = income per year
D_2 = 1 ; high school degree
0 ; others
D_3 = 1 ; college degree (S1)
0 ; others

Average spending based on education:

- less than high school : $\alpha_1 + \beta X$ (*base category*)
- high school : $\alpha_1 + \alpha_2 + \beta X$
- university/college (S1): $\alpha_1 + \alpha_3 + \beta X$

Notes: Reference group is less than high school. Why?
How do we choose a base category?

Model with Several Qualitative Variables

Salary = f (experience, sex, what faculty)

$Y = \alpha_1 + \alpha_2 D_2 + \alpha_3 D_3 + \beta X + u$

Y = salary / year

X = years of teaching

D_2 = 1 ; male professor

0 ; female professor

D_3 = 1 ; professor in Faculty of Economics

0 ; others

Model with Several Qualitative Variables

Salary = f (experience, sex, what faculty)

$Y = \alpha_1 + \alpha_2 D_2 + \alpha_3 D_3 + \beta X + u$

Y = salary / year

X = years of teaching

D_2 = 1 ; male professor
 0 ; female professor

D_3 = 1 ; professor in Faculty of Economics
 0 ; others

- Average salary of a female professor outside FE: $\alpha_1 + \beta X$
- Average salary of a male professor outside FE: $\alpha_1 + \alpha_2 + \beta X$
- Average salary of a female professor inside FE: $\alpha_1 + \alpha_3 + \beta X$
- Average salary of a male professor inside FE: $\alpha_1 + \alpha_2 + \alpha_3 + \beta X$

Comparing 2 regressions

Saving (Y) = α_1 + α_2 Income (X) + u

The above model indicates that saving and income do not behave differently across sample and time.

However, in reality, there is a possibility that the model behaves differently before and after a certain event. Let say, behavior of saving is different between prior and post an economic crisis.

How to accommodate this changing in saving behavior?

The following model can be used in accommodating a change.

Period I, before crisis: $Y_i = \alpha_1 + \alpha_2 X_i + u_i$; $i = 1, 2, \ldots, n$

Period II, after crisis: $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$; $i = n+1, n+2, \ldots, N$

Possibilities in comparing those two models:

Case 1: $\alpha_1 = \beta_1$ and $\alpha_2 = \beta_2$
Case 2: $\alpha_1 \neq \beta_1$ and $\alpha_2 = \beta_2$
Case 3: $\alpha_1 = \beta_1$ and $\alpha_2 \neq \beta_2$
Case 4: $\alpha_1 \neq \beta_1$ and $\alpha_2 \neq \beta_2$

Case 1 : both models are the same, no shift
Case 4 : both models are different
and there is a shift

Dummy variables can be used in addressing this type of change.

Comparing 2 regression with dummy variables

$$Y_i = \alpha_1 + \alpha_2 D_i + \beta_1 X_i + \beta_2 D_i X_i + u_i$$

D_i = 1 ; observation from period 1
0 ; observation from period 2

Based on this representation, average saving (Y) in period:

I : $Y_i = (\alpha_1 + \alpha_2) + (\beta_1 + \beta_2) X_i$

II : $Y_i = \alpha_1 + \beta_1 X_i$

How do we know that there is a shifting in the model?

- 1: If $\alpha_2 = 0$ and $\beta_2 = 0 \Rightarrow$ No shifting
- 2: If $\alpha_2 \neq 0$ and $\beta_2 = 0 \Rightarrow$ the same slope sama, different intercept
- 3: If $\alpha_2 = 0$ and $\beta_2 \neq 0 \Rightarrow$ the same Intercept, different slope
- 4: If $\alpha_2 \neq 0$ and $\beta_2 \neq 0 \Rightarrow$ both intercept and slope are different

Using Dummy Variable to Formulate a Piecewise linear regression

Modeling a bonus for excellent sales agents:

Rules:

1. Commission is proportional with sales
2. Bonus is given for an agent that over a target, X*.

Y: Bonus

X: size of sales achieved by an agent

X^* : sales target

Define a dummy, $D = 1$; if $X > X^*$

0 ; if $X \leq X^*$

The commission can be modeled as follows:

$$\text{Commission} = \alpha_1 + \beta_1 X \quad ; \text{for } X < X^*$$

$$\text{Commission} = \alpha_1 + \beta_1 X + \beta_2(X - X^*) \quad ; \text{for } X > X^*$$

Using dummy formulation:

$$\text{Commission} = \alpha_1 + \beta_1 X + \beta_2(X - X^*) D$$