Dummy Variables

Part 5

EE325

Introductory Econometrics

Revision Aug 2020

List of the topics to cover

Dummy variables

- Concept of turning a qualitative variable into a quantitative one.
- Interpretation of estimated coefficients.

Interaction terms

- Crossing a dummy variable with another variable.

Other applications

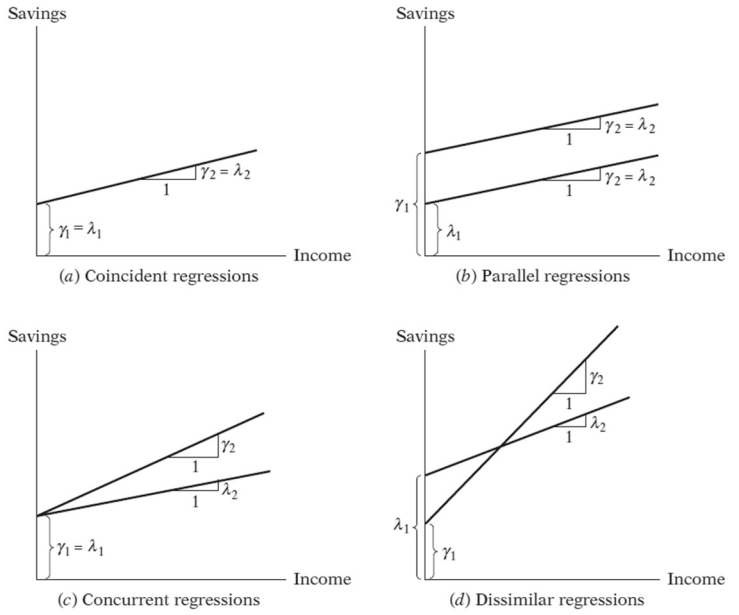
- Comparing dummy variable with the Chow Test.
- Another application: seasonal analysis.

Concept of ANOVA model

Recalling the Chow Test, a test for structural change, let's see all the possibilities from ex-ante and post crisis.

- $Y_t = \lambda_1 + \lambda_2 X_t + u_{1t}$ $\quad n_1 = 12$
- $Y_t = \gamma_1 + \gamma_2 X_t + u_{2t}$ $\quad n_2 = 14$
- $Y_t = \beta_1 + \beta_2 X_t + u_t$ $\quad n = (n_1 + n_2) = 26$

All possibilities of income-saving relationship



As discussed earlier, the major for overall test is that F-test is usually very general, when a null hypothesis is rejected.

Though a null hypothesis is rejected, we still do not know what and how, in this case, λ_1, γ_1 and λ_2, γ_2 are different. We would know if we keep nesting the F-test, which is way too much work.

If we can include a variable that separates ex-ante and post crisis period in a single equation, that would be ideal because we can see a difference, or no difference between pre and post crisis. We can also see its significance with a single t-test.

Not only that, if we can include another variable that can capture the slope for pre and post crisis, that is also very helpful since we can test its significance with a t-test.

Concept of ANOVA model

G. 277

So far, we have only dealt with continuous variables (weight, height, income, price, quantity, temperature, etc.) for both dependent and independent variables.

A natural problem arises since we know that there are so many real-world variables that is '**qualitative**'. A basic and most upfront example is gender. Consider our show size model.

- $shoesize_i = \beta_1 + \beta_2 gender_i + u_i$

We know for sure that gender is a very important factor determining shoe size. The problem is how can we incorporate this factor into this equation.

Gender is a binary variable (not chosen gender but biological), therefore there are only two possible encodings either

- $gender_i =$ male
- $gender_i =$ female

This model is called ANOVA model, or a model containing only quantitative variables or **dummy variable**.

Now let's consider the results.

Regression results

- $\hat{Y}_i = 42.375 - 4.25X_i$

$(0.5310) \quad (0.6503)$

$n = 24 \quad R^2 = 0.6600 \quad \bar{R}^2 = 0.6446$

$F(1,22) = 42.71 \qquad Prob > F = 0.000$

First, we look at how this result should be interpreted by considering the expected value. (For short, I go for shoe size variable as Y_i and the dummy variable by X_i)

- $E(Y_i|X_i = 0) =$

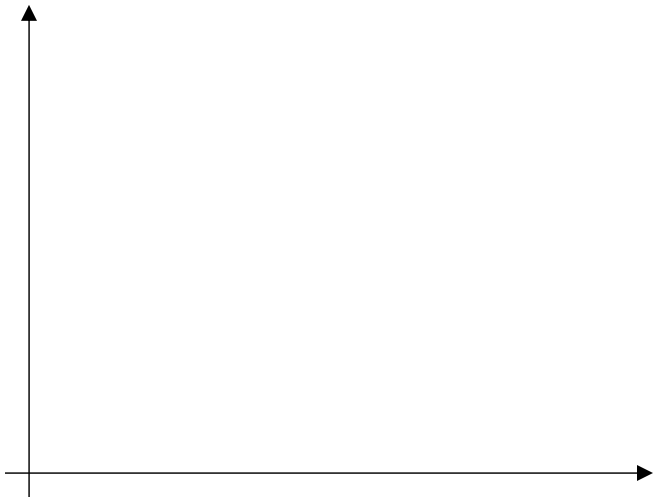
- $E(Y_i|X_i = 1) =$

(1) A caution: more-than-2-categories dummy variable

G. 281

If we plot each expected value, we see a difference between each group on this graph.

Expected value and coefficients



Data for the estimation are in this table. To distinguish gender, we only need one dummy variable to accommodate this difference. To be precise, we need only $n-1$ dummy variable(s) to incorporate n groups of the sample.

Shoe size	gender
42	0
40	1
39	1
36	1
40	0
39	1
36	1
38	1
37	1
39	1
39	1
44	0
37	1
43	0
41	1
36	1
45	0
39	1
42	0
38	1
38	1
41	0
38	1
42	0

Now let's assume that we will use chosen gender instead of biological gender, so we have 3 groups which are male, female, and LQBTQ+. Design the model below.

(2) A model with two dummy variables

G. 283

Now consider an ANOVA model with 2 dummy variables in the same model of show size. I created another false (ridiculous) dummy variable here as

- $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$

For gender variable (X_{2i}), encoding is the same (0 for male and 1 for female). For region (X_{3i}), assumed that 0 is living in Bangkok and 1 is living outside of Bangkok.

The interpretations of each category can be

- $E(Y_i|X_{2i} = 0, X_{3i} = 0) =$
- $E(Y_i|X_{2i} = 0, X_{3i} = 1) =$
- $E(Y_i|X_{2i} = 1, X_{3i} = 0) =$
- $E(Y_i|X_{2i} = 1, X_{3i} = 1) =$

(2) A model with two dummy variables

G. 283

Assumed that we have a fake result as follows,

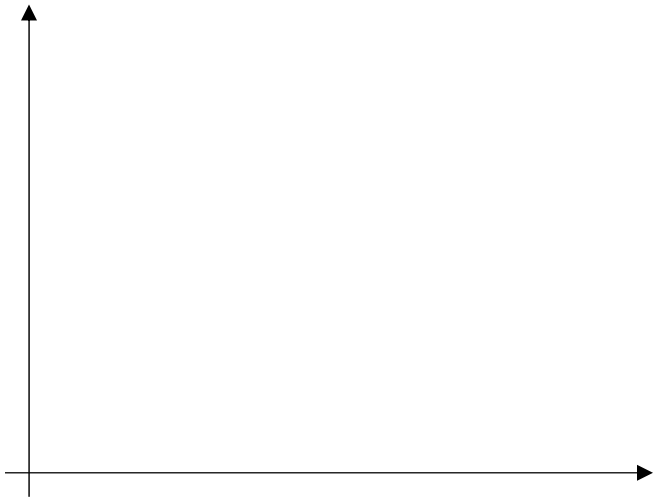
Regression results

- $\hat{Y}_i = 42.375 - 4.25X_{2i} + 1.1X_{3i}$

  $(0.5310)\quad (0.6503)\quad (2.2345)$

Therefore, if we plot the expected value of each group, it can be portrayed below.

Expected value and coefficients



In summary, for a model with 1 quantitative variable,

- $n-1$ dummy variables is to be included in the model to distinguish qualitative features of a variables, leading to n outcomes.

For a model with more than one qualitative variable,

- $n-1$ dummy variables is to be included in the model **for each feature**.
- Total outcome will be $2n$.

However, if we run a regression with STATA, we can encode many categories into one variable. There is a certain way that STATA can set up dummy variables without generating tons of them.

(3) A model with both quantitative and qualitative variable

G. 284

Regression models containing a mixture of both types of variables are called **ANCOVA models**. (Analysis of covariance)

Let's go back to our real sample of the show size model, now we include height (a quantitative continuous variable) and gender (a qualitative discrete variable) in this model as follows.

- $shoesize_i = \beta_1 + \beta_2 height_i + \beta_3 gender_i + u_i$

Reported results are in shown below.

Regression results

- $\hat{Y}_i = 12.0954 + 0.1739X_{2i} - 1.8589X_{3i}$

$(5.7413) \quad (0.0329) \quad (0.6285)$

$n = 24 \quad R^2 = 0.8541 \quad \bar{R}^2 = 0.8402$

$F(2,21) = 61.45 \quad Prob > F = 0.000$

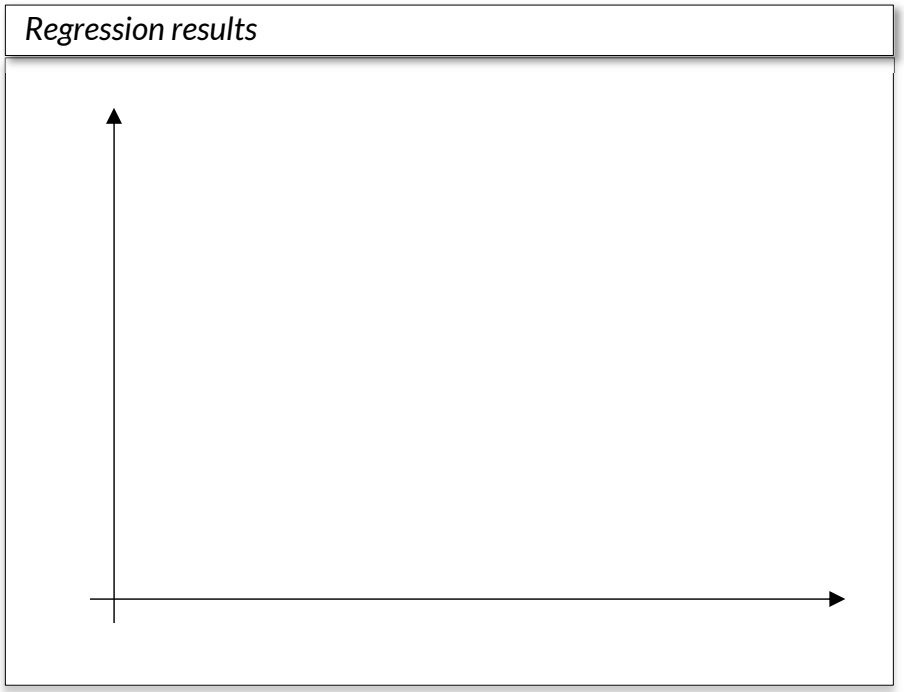
For this ANCOVA model, we have both quantitative and qualitative variables included. Each of them determines shoe size in a different way. Therefore, we need to interpret both height and gender that coexist in the same model.

Let's leave height for now sine we know it is the slope of the SRF, consider these outcomes of gender.

- $E(Y_i|X_{3i} = 0) =$
- $E(Y_i|X_{3i} = 1) =$

The best way to describe how this model works is to illustrate in the graph below.

Regression results



Crossing a dummy variable with another variable

G. 288

Dummy variables can be crossed to study differential effect from two or more dummies stacked together. Consider the following example of wage equation.

- $Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 (D_{2i} D_{3i}) + \beta_5 X_{5i} + u_i$

where Y_i is hourly wage ; X_{5i} is years of education

D_{2i} = 1 if female, 0 otherwise

D_{3i} = 1 if nonwhite and non-Hispanic, 0 otherwise

Regression results

- $\hat{Y}_i = -0.261 - 2.3606 D_{2i} - 1.7327 D_{3i}$

$$+2.1289 D_{2i} D_{3i} + 0.8028 X_{5i}$$

All possibilities are

- $E(Y_i | D_{2i} = 0; D_{3i} = 0) =$
- $E(Y_i | D_{2i} = 0; D_{3i} = 1) =$
- $E(Y_i | D_{2i} = 1; D_{3i} = 0) =$
- $E(Y_i | D_{2i} = 1; D_{3i} = 1) =$

$D_{2i} D_{3i}$ represents **additional effect**. This variable is called **interaction dummy**, effect of the two attributes considered individually.

Let's consider each effect here

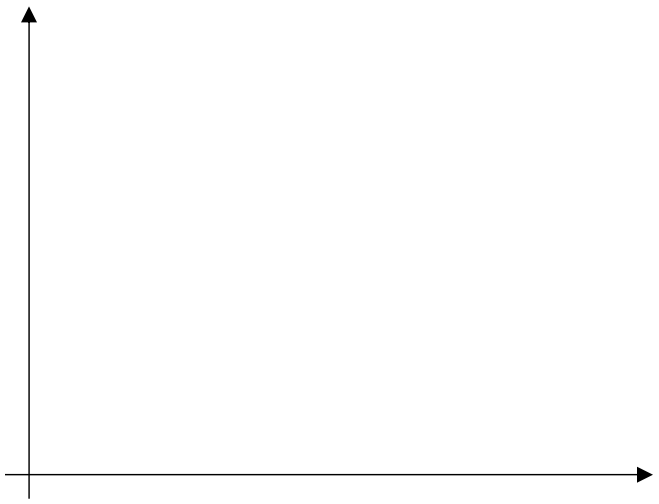
- Gender alone:
- Race alone:
- Gender and race:

We may see that if both gender and race affect wage per hour simultaneously, we can see some more social interpretations. Let's illustrate all the difference in a following plot.

Crossing a dummy variable with another variable

G. 290

A plot of an interaction dummy



If β_4 turns out to be significant, which it is, it means that gender and race have some kind of interaction effect.

In other words, a female, nonwhite and non-Hispanic worker is in between the effect of gender and race alone.

This is a difference that we cannot see when we only have two separate dummies, since the interpretation for each of them holds another constant.

We can also cross a dummy with a continuous variable. Let's go back to the Chow Test with pre and post crisis data.

Year	Savings	Income	Dum
1970	61	727.1	0
1971	68.6	790.2	0
1972	63.6	855.3	0
1973	89.6	965	0
1974	97.6	1054.2	0
1975	104.4	1159.2	0
1976	96.4	1273	0
1977	92.5	1401.4	0
1978	112.6	1580.1	0
1979	130.1	1769.5	0
1980	161.8	1973.3	0
1981	199.1	2200.2	0
1982	205.5	2347.3	1
1983	167	2522.4	1
1984	235.7	2810	1
1985	206.2	3002	1
1986	196.5	3187.6	1
1987	168.4	3363.1	1
1988	189.1	3640.8	1
1989	187.8	3894.5	1
1990	208.7	4166.8	1
1991	246.4	4343.7	1
1992	272.6	4613.7	1
1993	214.4	4790.2	1
1994	189.4	5021.7	1
1995	249.3	5320.8	1

Crossing a dummy variable with another variable

G. 286

If we only include pre and post dummy variable into the equation, we can only see intercept difference. Therefore, we need to add another interaction term into this model to see if there is slope difference between two periods or not.

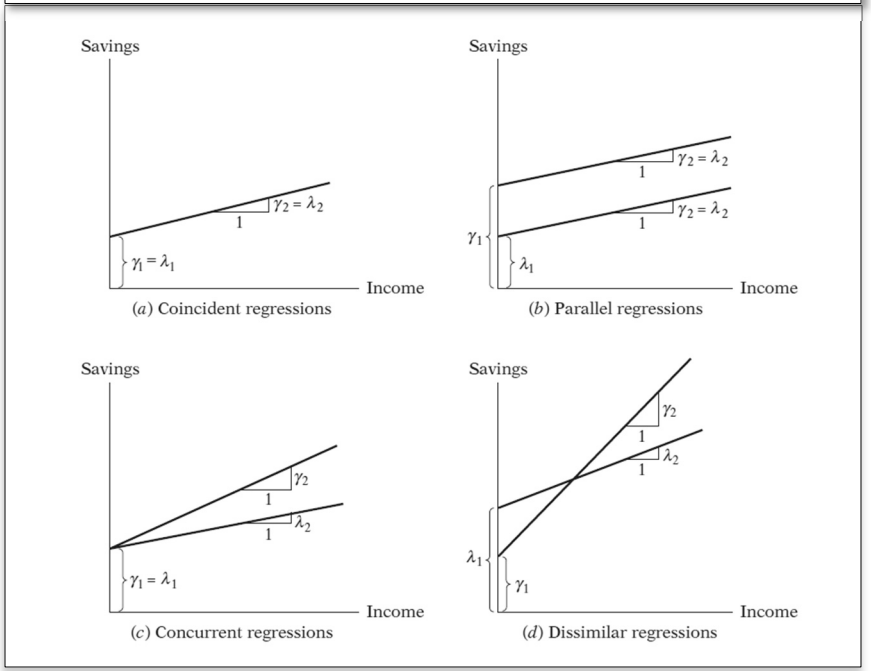
- $Y_t = \beta_1 + \beta_2 D_t + \beta_3 X_t + \beta_4 (D_t X_t) + u_t$

Consider all the possibilities from this model.

- $E(Y_t | D_t = 0; X_t = 0) =$
- $E(Y_t | D_t = 0; X_t \neq 0) =$
- $E(Y_t | D_t = 1; X_t = 0) =$
- $E(Y_t | D_t = 1; X_t \neq 0) =$

Each case does not represent each scenario from the illustration on the right-hand side directly.

All possibilities of income-saving relationship



Therefore,

- β_1 represents
- β_2 represents
- β_3 represents
- β_4 represents

Seasonal analysis

G. 290

Dummy variables can be used in multiple scenarios that we want to 'break' our data. Consider a model here with dummy variable that take seasonal breaks into account.

- $Y_t = \beta_1 D_{1t} + \beta_2 D_{2t} + \beta_3 D_{3t} + \beta_4 D_{4t} + u_t$

where Y_t is sales of refrigerator

D_{1t} = 1 if first quarter, 0 otherwise

D_{2t} = 1 if second quarter, 0 otherwise

D_{3t} = 1 if third quarter, 0 otherwise

D_{4t} = 1 if fourth quarter, 0 otherwise

Notice that we leave out the intercept here, as usually the intercept term will represent base case. This model is very easy to interpret since a data point can only fits within a quarter.

Each coefficient represents an average sale figure of each quarter. All of them are significantly different from zero.

Let's consider another model that includes the intercept, but now we have to drop one of the dummies as usual.

Regression results

- $\hat{Y}_t = 1{,}222.125 D_{1t} + 1{,}467.5 D_{2t} + 1{,}569.75 D_{3t} + 1{,}160 D_{4t}$

- $Y_t = \beta_1 + \beta_2 D_{2t} + \beta_3 D_{3t} + \beta_4 D_{4t} + u_t$

where Y_t is sales of refrigerator

D_{2t} = 1 if second quarter, 0 otherwise

D_{3t} = 1 if third quarter, 0 otherwise

D_{4t} = 1 if fourth quarter, 0 otherwise

The results and interpretation will be different since the β_1 will be our base case, or first quarter sales. Values of the dummy coefficients are different since the interpretation now is a comparison between the base case versus each quarter.

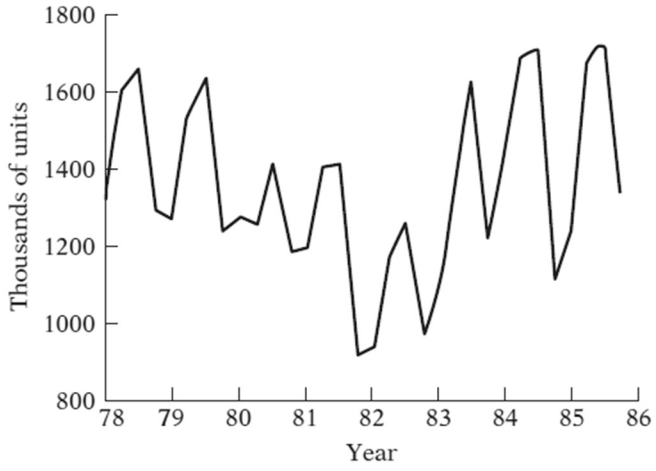
Regression results

- $\hat{Y}_t = 1{,}222.125 + 245.375 D_{2t} + 347.625 D_{3t} - 62.125 D_{4t}$

Seasonal analysis

G. 293

Seasonal variation



Significance of all coefficients in both model must stay align, though the standard errors are different due to interpretation of coefficients.

However, we do not take the problem of autocorrelation into account, which is very likely in this case (when $cov(u_i, u_j) \neq 0$). This is a time-series data in which sales number does not purely base on season, but also other built-up variables, such as GDP.

We will consider this problem in the next chapter.