Chapter 5

Dummy variable

A problem of Chow test

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

› $Y_t = \lambda_1 + \lambda_2 X_t + u_{1t}$ $\quad n_1 = 12$

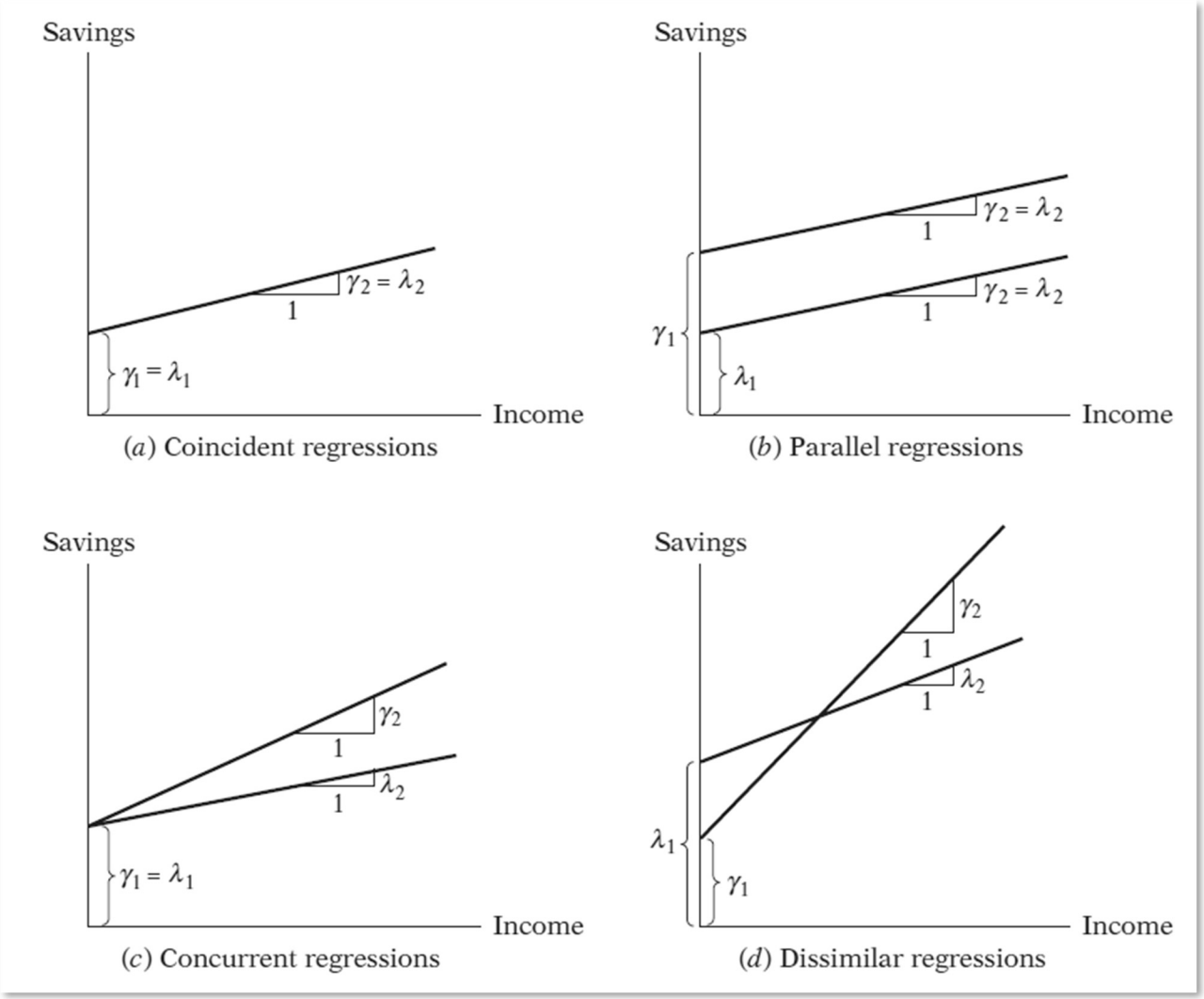
› $Y_t = \gamma_1 + \gamma_2 X_t + u_{2t}$ $\quad n_2 = 14$

› $Y_t = \beta_1 + \beta_2 X_t + u_t$ $\quad n = (n_1 + n_2) = 26$

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.

A problem of Chow test



A problem of Chow test

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 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 as well.

We are gradually going to implement this concept step-by-step.

(1) ANOVA model

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.

› $ss_i = \beta_1 + \beta_2 sex_i + u_i$

where ss_i is shoe size and sex_i is a binary variable, therefore there are only two possible encodings either

› sex_i = otherwise

› sex_i = female

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

(1) ANOVA model

Given that the result of this regression model is

› $\widehat{ss}_i = 41.833 - 3.583 sex_i$

First, we look at how this result should be interpreted by considering the expected value. Note that we encode $sex_i = 0$ for otherwise and $sex_i = 1$ for female

› $E(\widehat{ss}_i | sex_i = 0) =$

› $E(\widehat{ss}_i | sex_i = 1) =$

(1) ANOVA model

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

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.

(2) Three categories

Now let's assume that we will use chosen gender instead of biological sex, so we have 3 groups which are male, female, and LQBTQ+. Only one option can be chosen among these. Given that

› $D_{2i} = 0$ for otherwise ; $D_{2i} = 1$ for female

› $D_{3i} = 0$ for otherwise ; $D_{3i} = 1$ for LGBTQ+

The estimated model becomes

› $\widehat{ss}_i = \hat{\beta}_1 + \hat{\beta}_2 D_{2i} + \hat{\beta}_3 D_{3i}$

(2) Three categories

Find the expected value and interpretation for

› $E(\widehat{ss}_i | D_{2i} = 0; D_{3i} = 0) =$

› $E(\widehat{ss}_i | D_{2i} = 0; D_{3i} = 1) =$

› $E(\widehat{ss}_i | D_{2i} = 1; D_{3i} = 0) =$

› $E(\widehat{ss}_i | D_{2i} = 1; D_{3i} = 1) =$

(2) Three categories

Given that the result of this regression model is

› $\widehat{ss}_i = 41.833 - 2.751D_{2i} + 0.63D_{3i}$

Plot each group onto this graph.

(3) Two dummy variables

Let's say we now have 2 quantitative variables, sex and area where

› $D_{2i} = 0$ for otherwise ; $D_{2i} = 1$ for female

› $D_{3i} = 0$ for otherwise ; $D_{3i} = 1$ for Bangkokian

The estimated model becomes

› $\widehat{ss}_i = \hat{\beta}_1 + \hat{\beta}_2 D_{2i} + \hat{\beta}_3 D_{3i}$

(3) Two dummy variables

Find the expected value and interpretation for

› $E(\widehat{ss}_i | D_{2i} = 0; D_{3i} = 0) =$

› $E(\widehat{ss}_i | D_{2i} = 0; D_{3i} = 1) =$

› $E(\widehat{ss}_i | D_{2i} = 1; D_{3i} = 0) =$

› $E(\widehat{ss}_i | D_{2i} = 1; D_{3i} = 1) =$

(3) Two dummy variables

Given that the result of this regression model is

› $\widehat{ss}_i = 41.833 - 3.182D_{2i} + 1.54D_{3i}$

Plot each group onto this graph.

(4) ANCOVA model

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.

› $ss_i = \beta_1 + \beta_2 hei_i + \beta_3 sex_i + u_i$

where ss_i is shoe size, hei_i is height and sex_i is a binary variable as usual.

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.

(4) ANCOVA model

Given that the result of this regression model is

› $\widehat{ss}_i = 12.017 + 0.172hei_i - 1.456sex_i$

Now consider each case when

› $E(\widehat{ss}_i|sex_i = 0) =$

› $E(\widehat{ss}_i|sex_i = 1) =$

(4) ANCOVA model

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

(1) Dummy and dummy

Dummy variables can be crossed to study differential effect from two or more dummies stacked together. Consider the following example of the same equation, instead we add a cross-product term here.

› $ss_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 (D_{2i} D_{3i}) + u_i$

where

› $D_{2i} = 0$ for otherwise ; $D_{2i} = 1$ for female

› $D_{3i} = 0$ for otherwise ; $D_{3i} = 1$ for Bangkokian

β_4 represents **additional effect**. The term is called **interaction dummy**, effect of the two attributes considered individually.

Therefore, when both dummies or either D_{2i} or D_{3i} is zero, β_4 will be zero. This coefficient can be tested only the case of **Bangkokian female** if there is any additional significant effect.

(2) Dummy and continuous variable

A dummy variable can be crossed with another continuous variable as well. Given that

› $ss_i = \beta_1 + \beta_2 hei_i + \beta_3 sex_i + \beta_4 (hei_i sex_i) + u_i$

Let's find the expected value from estimated model.

› $E(\widehat{ss}_i | sex_i = 0) =$

› $E(\widehat{ss}_i | sex_i = 1) =$

(3) Example

Now all of the concepts are explained, let's take a look at our example from the estimation results. We consider the basic models.

1st model: › $ss_i = \beta_1 + \beta_2 hei_i + u_i$ all observation : n=66

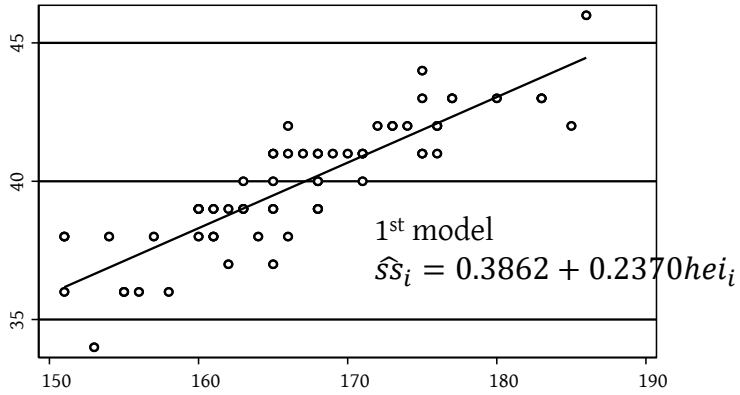
2nd model: › $ss_i = \beta_1 + \beta_2 hei_i + u_i$ male observation only : n=30

3rd model: › $ss_i = \beta_1 + \beta_2 hei_i + u_i$ female observation only : n=36

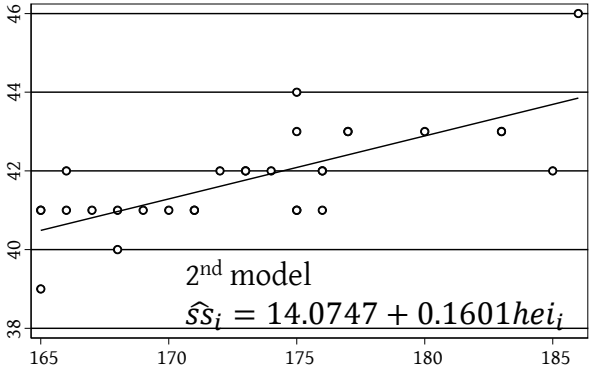
The results are listed here.

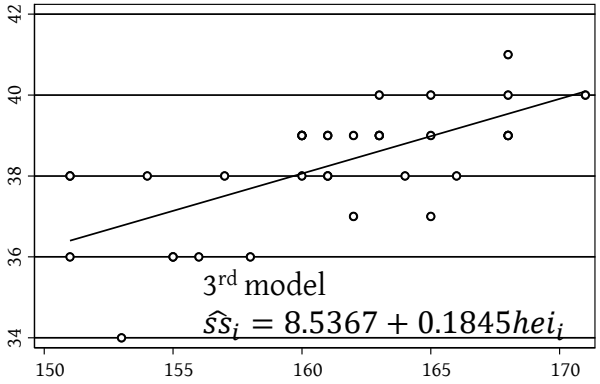
Model		Coefficient	P-value	R^2	$\bar{R}^2$
1st model	$\hat{\beta}_1$	0.3862	0.890	0.7598	0.7561
	$\hat{\beta}_2$	0.2370	0.000		
2nd model	$\hat{\beta}_1$	14.0747	0.008	0.5329	0.5162
	$\hat{\beta}_2$	0.1601	0.000		
3rd model	$\hat{\beta}_1$	8.5367	0.140	0.4487	0.4324
	$\hat{\beta}_2$	0.1845	0.000		

(3) Example



Plotting all the fitted regression line and scatter plot here.





(3) Example

To test if there is any difference between sex, we introduce two more models, incorporating a dummy and an interaction term.

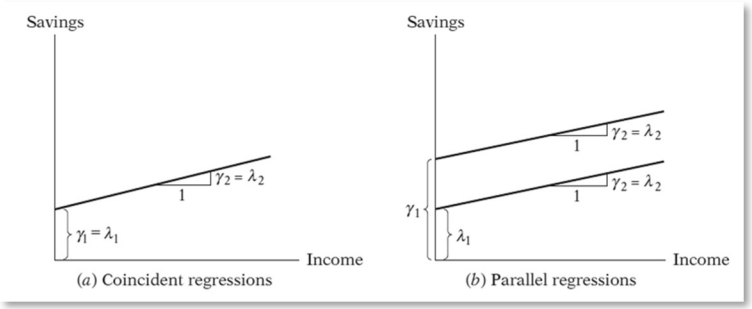
4th model: › $ss_i = \beta_1 + \beta_2 hei_i + \beta_3 sex_i + u_i$

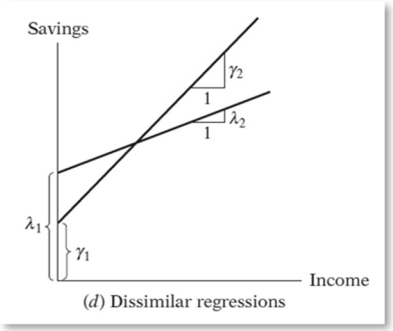
5th model: › $ss_i = \beta_1 + \beta_2 hei_i + \beta_3 sex_i + \beta_4 (sex_i hei_i) + u_i$

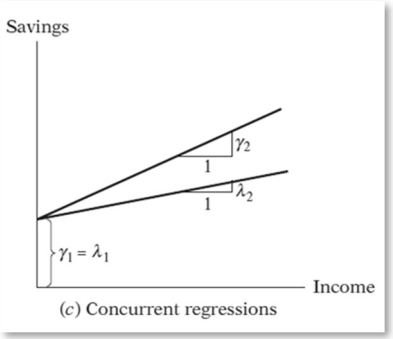
The results are listed here.

Model		Coefficient	P-value	R^2	$\bar{R}^2$
4th model	$\hat{\beta}_1$	12.0167	0.003	0.8061	0.8000
	$\hat{\beta}_2$	0.1720	0.000		
	$\hat{\beta}_3$	-1.460	0.000		
5th model	$\hat{\beta}_1$	14.0747	0.013	0.8070	0.7977
	$\hat{\beta}_2$	0.1601	0.000		
	$\hat{\beta}_3$	-5.5380	0.468		
	$\hat{\beta}_4$	0.0244	0.592		

(3) Example







In the 4th model, the dummy is used to test that the intercept of two groups are significantly different or not.

Meanwhile, in the 5th model we test that **both** the intercepts and slopes are significantly different **simultaneously** or not.
It turns out that they do not.

We can create another model that only test the slopes difference. This will be named as the 6th model.

Note: the pictures' purpose is only to illustrate the difference.

(3) Example

6th model: › $ss_i = \beta_1 + \beta_2 hei_i + \beta_3 (sex_i hei_i) + u_i$

The result is in the table below

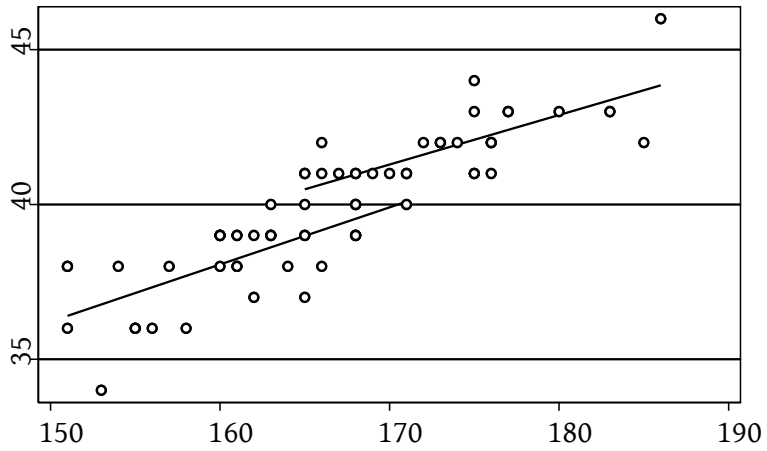
Model		Coefficient	P-value	R^2	$\overline{R}^2$
	$\hat{\beta}_1$	11.1802	0.004		
6th model	$\hat{\beta}_2$	0.1768	0.000	0.8054	0.7992
	$\hat{\beta}_3$	-0.0086	0.000		

This means that for female, the slope is a little bit less steep.

Now in this model, we can see that all the coefficients are statistically significant, as in the 4th model. The question is which model that we rely on.

The first thing to notice is that the 4th model has the highest value of $\overline{R}^2$.

(3) Example



This second thing, which is a lot more important, is that realistically, male and female should have different starting point and height should not affect them differently.

There is no reason to believe that the 6th model is true since there might hardly be any evidence suggesting alternating height affects shoe size differently. Male and female shoe size that are proportionally different makes a lot more sense.

See this plot to further elaborate my argument.