L6: Dummy variable regression models

Feng Li
feng.li@cufe.edu.cn

School of Statistics and Mathematics
Central University of Finance and Economics

Introduction

↛ Example–the data

TABLE 9.1 AVERAGE SALARY OF PUBLIC SCHOOL TEACHERS, BY STATE, 1986

Salary	Spending	D_2	D_3	Salary	Spending	D_2	D_3
19,583	3346	1	0	22,795	3366	0	1
20,263	3114	1	0	21,570	2920	0	1
20,325	3554	1	0	22,080	2980	0	1
26,800	4642	1	0	22,250	3731	0	1
29,470	4669	1	0	20,940	2853	0	1
26,610	4888	1	0	21,800	2533	0	1
30,678	5710	1	0	22,934	2729	0	1
27,170	5536	1	0	18,443	2305	0	1
25,853	4168	1	0	19,538	2642	0	1
24,500	3547	1	0	20,460	3124	0	1
24,274	3159	1	0	21,419	2752	0	1
27,170	3621	1	0	25,160	3429	0	1
30,168	3782	1	0	22,482	3947	0	0
26,525	4247	1	0	20,969	2509	0	0
27,360	3982	1	0	27,224	5440	0	0
21,690	3568	1	0	25,892	4042	0	0
21,974	3155	1	0	22,644	3402	0	0
20,816	3059	1	0	24,640	2829	0	0
18,095	2967	1	0	22,341	2297	0	0
20,939	3285	1	0	25,610	2932	0	0
22,644	3914	1	0	26,015	3705	0	0
24,624	4517	0	1	25,788	4123	0	0
27,186	4349	0	1	29,132	3608	0	0
33,990	5020	0	1	41,480	8349	0	0
23,382	3594	0	1	25,845	3766	0	0
20,627	2821	0	1				

Note: $D_2 = 1$ for states in the Northeast and North Central; 0 otherwise.
$D_3 = 1$ for states in the South; 0 otherwise.
Source: National Educational Association, as reported by *Albuquerque Tribune*, Nov. 7, 1986.

Introduction
Example – the question

- Does the average annual salary among (1) Northeast and North Center, (2) South and (3) West differ?
- Consider the model

$$Salary_i = \alpha + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i$$

where

$$D_{2i} = \begin{cases} 1 & \text{if ith state from (1)} \\ 0 & \text{otherwise} \end{cases}, D_{3i} = \begin{cases} 1 & \text{if ith state from (2)} \\ 0 & \text{otherwise} \end{cases}$$

- Then treat D_2 and D_3 as ordinary variables and do the linear regression as usual, simple!
- This is called **dummy variable regression**.

↪ Example – the interpretation

- With above data and model, we have the following result

$$Salary_i = 48014 + 1524 D_{2i} - 1721 D_{3i}$$

- The mean salary of teachers from (3) is \$48014. why?
- Teacher salary from (1) is \$1524 higher than the mean salary from (3).
- Teacher salary from (2) is \$1721 lower than the mean salary from (3).
- Interpretations for other quantities follow the way of linear regression model.

Introduction

↛ Example – dummy variable models without intercept

- If we cream a new dummy variable

$$D_{1i} = \begin{cases} 1 & \text{if ith state from (3)} \\ 0 & \text{otherwise} \end{cases}$$

- Then make regression model with D_1, D_2, D_3 without intercept, i.e.

$$Salary_i = \gamma_1 D_{1i} + \gamma_2 D_{2i} + \gamma_3 D_{3i} + \epsilon_i$$

- It interpreted as
  - γ_1: average salary from (3),
  - γ_2: average salary from (1),
  - γ_3: average salary from (2).
- What will $\hat{\gamma}_1$, $\hat{\gamma}_2$, $\hat{\gamma}_3$ be?
  - $\hat{\gamma}_1 =$ mean salary of teachers from $(3) = \hat{\alpha} = 48014$
  - $\hat{\gamma}_2 =$ mean salary of teachers from $(1) = \hat{\alpha} + \hat{\beta}_2 = 48014 + 1524 = 49538$
  - $\hat{\gamma}_3 =$ mean salary of teachers from $(2) = \hat{\alpha} + \hat{\beta}_3 = 48014 - 1721 = 46293$

- If you make a dummy regression with the above data as

$$Salary_i = \delta_0 + \delta_1 D_{1i} + \delta_2 D_{2i} + \epsilon_i$$

- what will $\hat{\delta}_0$, $\hat{\delta}_1$, $\hat{\delta}_2$ be?

The caution

- If the qualitative variable has m categories, introduce only $m-1$ dummy variables if the intercept is also included; need m dummies if intercept is not included. *What if you don't ? – Multicollinearity problem(in next lecture)*
  - Set up a model with dummies $D_1,\ D_2,\ ,...D_{m-1}$ is essential equivalent as that with dummies for any other combinations, e.g. $D_2,\ D_3,\ ,...D_m$.
- You don't always have to use 0 and 1 to indicate dummies, you can use any others, like

$$D_{1i} = \begin{cases} 2 & \text{if ith state from (3)} \\ 1 & \text{otherwise} \end{cases} \text{ or } D_{1i} = \begin{cases} 1 & \text{if ith state from (3)} \\ -1 & \text{otherwise} \end{cases}$$

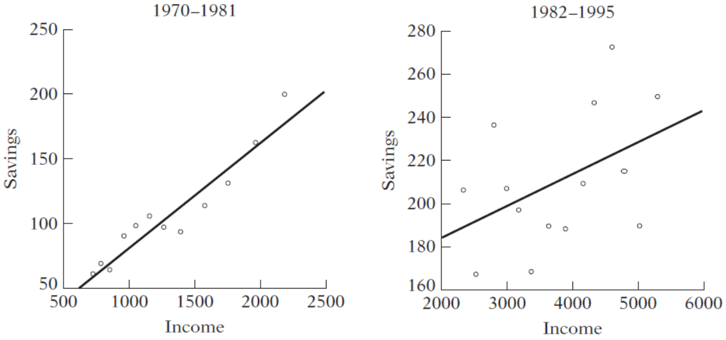
How to interpret it then? – see Exercise 9.5

Use dummy variable as an alternative to the Chow Test

- The Chow Test – review
  - The Chow Test is used to check if there is structural change in the dataset.
  - The null hypothesis: there is no structural change.
  - The test statistic is

$$F = \frac{(RSS_R - RSS_{UR})/k}{RSS_{UR}/(n_1 + n_2 - 2k)} \sim F(k, n_1 + n_2 - 2k)$$

- Look at this example (p.255)



- We want to check if there there is structural change in the two time period.

Use dummy variable as an alternative to the Chow Test (2)

- We can simply make a regression with dummies like

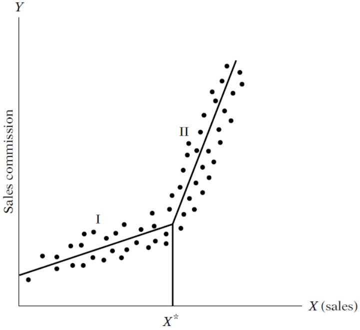
$$Savings_t = \alpha_1 + \alpha_2 D_t + \beta_1 \mathrm{Income}_t + \beta_2(D_t \mathrm{Income}_t) + u_t \quad \text{where}$$

$$D_{ti} = \begin{cases} 1 & \text{if ith obs. from 1982–1995} \\ 0 & \text{otherwise} \end{cases}$$

- The usual Chow Test can only show if there is a change or not, but
- $\hat{\beta}_2$ and $\hat{\beta}_2$ will show how much structure changed in the two period.

Use dummy variable models for piecewise linear regression

- Assume we have the following data



- A straight line will no fit it well. It is better to fit it with two lines ,

$$
\begin{aligned}
Y_i =& \alpha_1 + \alpha_2 X_i + u_i, \text{when } X_i < X^* \\
Y_i =& \beta_1 + \beta_2 X_i + u_i, \text{otherwise}
\end{aligned}
$$

- We can fit them together with the model

$$
Y_i = \alpha_1 + \beta_1 X_i + \beta_2 (X_i - X^*) D_i + u_i, \text{ where } D_i = \begin{cases} 1 & \text{if } X_i > X^* \\ 0 & \text{otherwise} \end{cases}
$$

Take home questions

- Assume you have a three-piecewise regression, how do you set up the dummies and how do you detect the right location of x^* as in previous slide?