Chapter 2 Multiple Regression
(Part 5)

1 Overview and Dummy Variable

- qualitative predictor (also called categorical variable)
- How to allocate codes (values) to qualitative predictor
- Interaction between quantitative and qualitative predictors
- Comparison of regression functions

An example: the insurance firm

In the **(Data)**: Y - speed of innovation, X_1 – size of a insurance firm, X_2 – type of firm: stock company or mutual company. Predictor variable X_2 is qualitative or categorical. It is obvious we cannot use model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ because X_2 is not real value.

- quantify (assign value to) a qualitative variable

$$X_2 \mapsto D = \{ \begin{array}{ll} 1, & \text{if stock company} \\ 0, & \text{otherwise} \end{array}$$

  D is called **Dummy variables**

- Then we can consider $Y = \beta_0 + \beta_1 X_1 + \beta_2 D + \varepsilon_i$

2 Interpretation of regression coefficients

- Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 D + \varepsilon_i$

  where $X_1 =$ size of firm, $\quad D = \{ \begin{array}{cc} 1 & \text{if stock company} \\ 0 & \text{if mutual company} \end{array}$

- $E\{Y\} = \beta_0 + \beta_1 X_1$ mutual firms (corresponding to $D = 0$)

  $E\{Y\} = \beta_0 + \beta_2 + \beta_1 X_1$ stock firms (corresponding to $D = 0$)

- β_2 measures the difference between mutual firms and stock firms

Insurance innovation example

$$\begin{array}{ccccccc} \hat{Y} & = & 33.87407 & - & 0.10174X_1 & + & 8.05547D \\ \text{(S.E.)} & & (1.81386) & & (0.00889) & & (1.45911) \end{array}$$

to check whether there is difference in the intercepts, we need to test

$$H_0 : \beta_2 = 0 \ \text{ vs } \ H_a : \beta_2 \neq 0:$$

Because

$$|t^*| = |\frac{8.05547 - 0}{1.45911}| = 5.52 \geq t(0.975; 17) = 2.110,$$

we conclude H_a: the intercepts are different from one another significantly

3 Qualitative predictor with more than two classes

- Consider an example. Response Y–tool wear, Predictor X_1 – tool speed; and predictor X_2–tool model with four classes: M1, M2, M3, M4. X_2 is qualitative.

- Quantify the qualitative predictor:

  Note that with two classes, we need $2 - 1 = 1$ variable, with four classes, we need $4 - 1 = 3$ variables

$$D_1 = \{ \begin{array}{ll} 1, & \text{if tool model M1} \\ 0, & \text{otherwise} \end{array}$$

$$D_2 = \{ \begin{array}{ll} 1, & \text{if tool model M2} \\ 0, & \text{otherwise} \end{array}$$

$$D_3 = \{ \begin{array}{ll} 1, & \text{if tool model M3} \\ 0, & \text{otherwise} \end{array}$$

  D_1, D_2, D_3 are dummy variables

  Thus, we have the following correspondence

X_1	$\leftrightarrow$	D_1	D_2	D_3
M1	$\leftrightarrow$	1	0	0
M2	$\leftrightarrow$	0	1	0
M3	$\leftrightarrow$	0	0	1
M4	$\leftrightarrow$	0	0	0

- Generally speaking, if a qualitative predictor has m classes ,we need $m-1$ dummy variables

- why dont we use m dummy variables? We can, but we need to drop the intercept. For the insurance firm data

$$D_1 = \{ \begin{array}{cc} 1 & \text{if stock company} \\ 0 & \text{if mutual company} \end{array}$$

$$D_2 = \{ \begin{array}{cc} 0 & \text{if stock company} \\ 1 & \text{if mutual company} \end{array}$$

Then the model

$$Y = \beta_1 X_1 + \beta_2 D_1 + \beta_3 D_2 + \varepsilon$$

$$E\{Y|\text{stock company}\} = \beta_2 + \beta_1 X_1$$

$$E\{Y|\text{mutual company}\} = \beta_3 + \beta_1 X_1$$

(IF we dont drop the intercept term, them the inverse $(X'X)^{-1}$ does not exist, because in

$$X = \begin{pmatrix} 1 & X_{11} & D_{11} & D_{12} \\ 1 & X_{11} & D_{11} & D_{12} \\ ... & & & \\ 1 & X_{n1} & D_{n1} & D_{n2} \end{pmatrix}$$

the summation of last two columns is the first column.)

interpretation of qualitative predictor with more than two classes

- For the tool wear example, its first-order model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 D_1 + \beta_3 D_2 + \beta_4 D_3 + \varepsilon$

- $E\{Y\} = \beta_0 + \beta_1 X_1$ tool model M4 (for $D_1 = 0, D_2 = 0, D_3 = 0$)

  $E\{Y\} = \beta_0 + \beta_2 + \beta_1 X_1$ tool model M1 (for $D_1 = 1, D_2 = 0, D_3 = 0$)

  $E\{Y\} = \beta_0 + \beta_3 + \beta_1 X_1$ tool model M2 (for $D_1 = 0, D_2 = 1, D_3 = 0$)

  $E\{Y\} = \beta_0 + \beta_4 + \beta_1 X_1$ tool model M3 (for $D_1 = 0, D_2 = 0, D_3 = 1$)

- Interpretation of regression coefficient:

  β_2: difference between M1 and M4. (Question: how to test whether the intercepts in the models for M1 and M4 are the same?)

  $\beta_2 - \beta_3$: difference between M1 and M2 (Question: how to test whether the intercepts in the models for M1 and M2 are the same?)

3.1 Another example

Why cannot we use 1,2,3, ... to denote the categorical variables

- Qualitative predictor: frequency of product use three classes: frequent user, occasional user and nonuser

- allocate codes X_1 by $\tilde{D}$,

class	$\tilde{D}$
frequent user	3
occasional user	2
nonuser	1

- $Y = \beta_0 + \beta_1 \tilde{D} + \varepsilon$

class	$E\{Y\}$
frequent user	$E\{Y\} = \beta_0 + 3\beta_1$
occasional user	$E\{Y\} = \beta_0 + 2\beta_1$
nonuser	$E\{Y\} = \beta_0 + \beta_1$

- Key implication and limitation:

$$E\{Y|\text{frequent user}\} - E\{Y|\text{occasional user}\}$$
$$= E\{Y|\text{occasional user}\} - E\{Y|\text{nonuser}\}$$

Thus, this allocation of code implies something inappropriate.

Indicator variables for quantitative variables

Sometimes it is even useful to use qualitative variables to represent quantitative variables after grouping. For example, when we consider 'age'

- group ages into four classes: under 21, 21-34, 35-39, above 40. Then, 'age' becomes qualitative predictor, and we need three qualitative predictors (but lose three degrees of freedom)

- advantage: no need to check linearity of regression function

- it is recommended in a large-scale study when the loss of several degrees of freedom is not much

Other codes for qualitative variables

- Consider the insurance firms again: Y - speed of innovation, X_1 size of firm,

$$X_2 \mapsto D = \{ \begin{array}{cc} 1 & \text{if stock company} \\ 0 & \text{if mutual company} \end{array}$$

- alternative code for X_2:

$$X_2 \mapsto D = \{ \begin{array}{cc} 1 & \text{if stock company} \\ -1 & \text{if mutual company} \end{array}$$

  for stock company: $E\{Y\} = \beta_0 + \beta_2 + \beta_1 X_1$ (for $D = 1$)

  for mutual company: $E\{Y\} = \beta_0 - \beta_2 + \beta_1 X_1$ (for $D = -1$)

3.2 Interaction between quantitative and qualitative predictors

- Consider the insurance data again,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 D + \beta_3 D * X_1 + \varepsilon$$

  $X_1 =$ is the size of firm and

$$D = \{ \begin{array}{cc} 1 & \text{if stock company} \\ 0 & \text{if mutual company} \end{array}$$

- different regression coefficients

  regression model for stock firms: $\beta_0 + \beta_2 + (\beta_1 + \beta_3) X_1$

  regression model for mutual firm: $\beta_0 + \beta_1 X_1$

  β_2 is the difference of intercepts for two types of firms

  β_3 is the difference of regression effects/slope

3.3 More consideration and Comparison of models for different categories

Suppose we have response Y (say son or daughter's height) with quantitative variables X_1, X_2 (Father's height and mother's height) and qualitative variable D, say $D = 1$ for Son, and $D = 0$ for daughter.

- If you believe there is no difference between son and daughter's height and dependence on their parents, then you may consider a general model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- IF you believe their heights differ mainly due of their gender then you may consider model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 D + \varepsilon$$

which is equivalent to

$$\begin{array}{rccccl} \text{Son:} & Y & = & (\beta_0 + \beta_3) & + & \beta_1 X_1 + \beta_2 X_2 + \varepsilon \\ \text{Daughter:} & Y & = & \beta_0 & + & \beta_1 X_1 + \beta_2 X_2 + \varepsilon \end{array}$$

Notice the common coefficients and different coefficients.

- If you believe parents' heights have different effect on son and daughter respectively, then consider

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 D * X_1 + \beta_5 D * X_2 + \varepsilon$$

which is equivalent to

$$\begin{array}{rccccccccl} \text{Son:} & Y & = & \beta_0 & + & (\beta_1 + \beta_4) X_1 & + & (\beta_2 + \beta_5) X_2 & +\varepsilon \\ \text{Daughter:} & Y & = & \beta_0 & + & \beta_1 X_1 & + & \beta_2 X_2 & +\varepsilon \end{array}$$

Notice the common coefficients and different coefficients.

- If you believe son and daughter height are completely different, then consider

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_3 X_1 + \beta_5 X_3 X_2 + \varepsilon$$

which is equivalent to

$$\begin{array}{rcccccccc} \text{Son:} & Y & = & (\beta_0 + \beta_3) & + & (\beta_1 + \beta_4) X_1 & + & (\beta_2 + \beta_5) X_2 & +\varepsilon \\ \text{Daughter:} & Y & = & \beta_0 & + & \beta_1 X_1 & + & \beta_2 X_2 & +\varepsilon \end{array}$$

They are actually two completely different models. They are equivalent to fit two completely models to two data (one for boys and another for girls).

- If you want to answer whether models for boys' height and girls' height are the same, it is equivalent to test $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$; if you want to test whether father's effect on both son's height and daughter's height are the same, you need to test $H_0 : \beta_4 = 0$; ....

Insurance innovation example

- For the general model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 D + \beta_3 D X_1 + \varepsilon$$

The estimated model is

$$\begin{array}{ccccccccc} \hat{Y} & = & 33.8383 & - & 0.1015X_1 & + & 8.13125D & - & 0.0004171D * X_1 \\ \text{(S.E.)} & & (2.44) & & (0.013) & & (3.65) & & (0.0183) \end{array}$$

- Test whether the effect of firm size change with firm type

$$H_0 : \beta_3 = 0, \ vs \ \ H_a : \beta_3 \neq 0$$

$$t^* = \frac{b_3}{s\{b_3\}} = \frac{-0.0004171}{0.01833} = -0.02,$$

as $|t^*| \leq t(0.975; 16) = 2.120$, conclude H_0 and can adopt the model with no interaction term, or the effect of firm size does not change significantly with firm type

4 more than one qualitative predictor

- Consider Y –advertising expenditure; X_1–sales; X_2–type of firm (incorporated, not incorporated); X_3–quality of sales management (high or low)
for X_2, introduce dummy variable $D_1 = \{ \begin{array}{cc} 1 & \text{if firm incorporated} \\ 0 & \text{otherwise} \end{array}$
for X_3, introduce dummy variable $D_2 = \{ \begin{array}{cc} 1 & \text{if quality of sales management high} \\ 0 & \text{otherwise} \end{array}$
- A model for the possible intercept difference

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 D_1 + \beta_3 D_2 + \varepsilon$$

- A model with certain interaction added

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 D_1 + \beta_3 D_2 + \beta_4 D_1 X_1 + \beta_5 D_2 X_1 + \beta_6 D_1 D_2 + \varepsilon$$

-

X_2	X_3	$E\{Y\}$
incorporated	high	$(\beta_0 + \beta_2 + \beta_3 + \beta_6) + (\beta_1 + \beta_4 + \beta_5)X_1$
not incorporated	high	$(\beta_0 + \beta_3) + (\beta_1 + \beta_5)X_1$
incorporated	low	$(\beta_0 + \beta_2) + (\beta_1 + \beta_4)X_1$
not incorporated	low	$\beta_0 + \beta_1 X_1$

- Question: why dont we consider the cross interaction between different dummy variables for one categorical predictor? [Because even if you do so, their product is 0]

4.1 Example 1: Soap production lines example

- Soap production lines example (see **(Data)**) Y - amount of scrap, X_1 - line speed, D - code for two possible production lines $D = 1$ or $D = 0$, 27 observations

- Full model $Y = \beta_0 + \beta_1 X_1 + \beta_2 D + \beta_3 X_1 D + \varepsilon$

- regression function for production 1: $(\beta_0 + \beta_2) + (\beta_1 + \beta_3)X$

  regression function for production 2: $\beta_0 + \beta_1 X$

Test the identity of two regression functions

To test whether two production lines have the same model

Reduced model $\quad Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$

Against, two models are different

Full model $\quad Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 D_i + \beta_3 X_{i1} D_i + \varepsilon_i$

AOVA for the full model (see **R-code**)

source	SS	df	MS
X1	149661	1	149661
D	18694	1	18694
DX1	810	1	810
residual	9904	23	431

we have

$$SSE(F) = 9904, \ (DF = 23)$$

and

$$\begin{aligned} SSR(D, DX_1|X_1) &= SSR(D|X_1) + SSR(DX_1|X_1, D) \\ &= 18694 + 810 = 19504, \ (DF = 21) \end{aligned}$$

- $H_0 : \beta_2 = \beta_3 = 0$

  H_a : not both $\beta_2 = 0$ and $\beta_3 = 0$

$$\begin{aligned} F^* &= \frac{SSR(D, DX_1|X_1)}{2} \div \frac{SSE(X_1, D, DX_1)}{27 - 4} \\ &= 22.65 \geq F(0.99, 2, 23) = 5.67 \end{aligned}$$

  conclude H_a, and regression functions for two lines are not identical

Test the equality of slopes of two regression functions

- $H_0 : \beta_3 = 0,\ H_a : \beta_3 \neq 0$
- $$\begin{aligned} F^* &= \frac{SSR(DX_1|X_1, D)}{1} \div \frac{SSE(X_1, D, DX_1)}{27-4} \\ &= 1.88 \leq F(0.99, 1, 23) = 7.88 \end{aligned}$$

  conclude H_0 and slopes for two regression functions are the same

4.2 Example 2: SENIC

The primary objective of the Study on the Efficacy of Nosocomial Infection Control (**SENIC** Project) was to determine whether infection surveillance and control programs have reduced the rates of nosocomial (hospital-acquired) infection in United States hospitals. This data set consists of a ramdom sample of 113 hospitals selected from the original 338 hospitals surveyed.

Each line of the data set has an identification number and provides information on 11 other variables for a single hospital. The data presented here are for the 1975-76 study period. The 12 variables are:

1 Identification number: 1-113

2 Length of stay: Average length of stay of all patients in hospital (in days)

3 Age: Average age of patients (in years)

4 Infection risk: Average estimated probability of acquiring infection in hospital (in percent)

5 Routine culturing ratio: Ratio of number of cultures performed to number of patients without signs or symptoms of hospital-acquired infection, times 100

6 Routine chest X-ray ratio: Ratio of number of X-rays performed to number of patients without signs or symptoms of pneumonia, times 100

7 Number of beds: Average number of beds in hospital during study period

8 Medical school affiliation: 1=Yes, 2=No

9 Region: Geographic region, where: 1=NE, 2=NC, 3=S, 4=W

10 Average daily census: Average number of patients in hospital per day during study period

11 Number of nurses: Average number of full-time equivalent registered and licensed practical nurses during study period (number full time plus one half the number part time)

12 Available facilities and services: Percent of 35 potential facilities and services that are provided by the hospital

For **(Data)**, consider a model of regressing infectious risk Y against age X_1, routine culturing ratio X_2, average daily census X_3, available facilities and service X_4, Medical school affiliation X_5. For each region, we have can find a model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon,$$

are the estimated regression functions similar for the four regions? Discuss.

Let $D_1 = 1$ if in region NE; otherwise 0;

Let $D_2 = 1$ if in region NC; otherwise 0;

Let $D_3 = 1$ if in region S; otherwise 0;

We consider full model

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \alpha_1 D_1 + \alpha_1 D_2 + \alpha_1 D_3 \\ &\quad + \beta_{11} D_1 X_1 + \beta_{12} D_1 X_2 + \beta_{13} D_1 X_3 + \beta_{14} D_1 X_4 + \beta_{15} D_1 X_5 \\ &\quad + \beta_{21} D_2 X_1 + \beta_{22} D_2 X_2 + \beta_{23} D_2 X_3 + \beta_{24} D_2 X_4 + \beta_{25} D_2 X_5 \\ &\quad + \beta_{31} D_3 X_1 + \beta_{32} D_3 X_2 + \beta_{23} D_3 X_3 + \beta_{34} D_3 X_4 + \beta_{35} D_3 X_5 \\ &\quad + \varepsilon, \end{aligned}$$

If there is no region effect, then the reduced model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon,$$

For the full model, we have

$$SSE(F) = 89.276 \quad DF = 89$$

For the reduced model, we have

$$SSE(R) = 110.933 \quad DF = 107$$

Thus

$$F^* = \frac{(110.933 - 89.276)/18}{89.276/89} = 1.1994 < F(0.95, 18, 89) = 1.73$$

Thus, we don't think different regions have different models. See the **R-code**.

5 Time series application

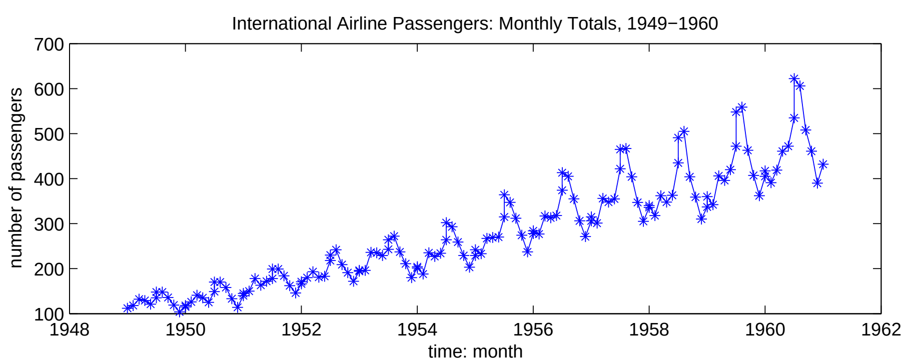
- For example, Y - quarterly sales, X_1 - quarterly advertisement expenditures, X_2 - quarterly personal disposable income.

$$D_1 = \begin{cases} 1, & \text{if first quarter} \\ 0, & \text{otherwise} \end{cases}, \quad D_2 = \begin{cases} 1 & \text{if second quarter} \\ 0 & \text{otherwise} \end{cases},$$

$$D_3 = \begin{cases} 1, & \text{if third quarter} \\ 0, & \text{otherwise} \end{cases}$$

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \beta_3 D_{t1} + \beta_4 D_{t2} + \beta_5 D_{t3} + \varepsilon_t$$

- another example



In this example, we can consider a model

$$Y_t = \beta_0 + \beta_1 * t + \beta_2 D_1 + ... + \beta_{11} D_{11} + \varepsilon_t$$

where $D_1, ..., D_{11}$ are dummy variables denoting the month of a year (how).