第 1 章
迴歸專題

—— 1.1 當變數進行線性轉換 ——

性質 1.1 線性轉換 (*linear transformation*)

首先利用原始資料 $\{(X_i, Y_i)\}_{i=1}^n$ 進行迴歸分析, 其 *OLS* 估計式分別爲 $\hat{\alpha}$ 與 $\hat{\beta}$, 即

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

給定 a, b, c, d 爲已知常數, 若將原始資料進行線性轉換, 即

$$X_i^* = aX_i + b$$
$$Y_i^* = cY_i + d$$

則利用線性轉換後的資料 $\{(X_i^*, Y_i^*)\}_{i=1}^n$ 進行迴歸分析, 其 *OLS* 估計式爲 $\hat{\alpha}^*$ 與 $\hat{\beta}^*$, 即

$$\hat{Y}_i^* = \hat{\alpha}^* + \hat{\beta}^* X_i^*$$

則利用原始資料與線性轉換過的資料之 *OLS* 估計式之間有以下關係:

$$\hat{\beta}^* = \frac{c}{a}\hat{\beta} \quad \text{(斜率項不受平移項 } b, d \text{ 的影響)}$$
$$\hat{\alpha}^* = c\,\hat{\alpha} + d - \frac{bc}{a}\hat{\beta}$$

《***Proof***》

首先得知 $\bar{X}^* = a\bar{X} + b$, $S_{X^*}^2 = a^2 S_X^2$, $\bar{Y}^* = c\bar{Y} + d$, $S_{Y^*}^2 = c^2 S_Y^2$, $S_{X^*Y^*} = acS_{XY}$。根據迴歸係數的公式可得到

$$\hat{\beta}^* = \frac{S_{X^*Y^*}}{S_{X^*}^2} = \frac{acS_{XY}}{a^2 S_X^2} = \frac{c}{a}\underbrace{\frac{S_{XY}}{S_X^2}}_{=\hat{\beta}} = \frac{c}{a}\hat{\beta}$$

$$\begin{aligned}\hat{\alpha}^* &= \bar{Y}^* - \hat{\beta}^*\bar{X}^* = c\bar{Y} + d - \frac{c}{a}\hat{\beta} \times (a\bar{X} + b) = c\bar{Y} - c\hat{\beta}\bar{X} + d - \frac{bc}{a}\hat{\beta} \\ &= c\underbrace{(\bar{Y} - \hat{\beta}\bar{X})}_{=\hat{\alpha}} + d - \frac{bc}{a}\hat{\beta} = c\,\hat{\alpha} + d - \frac{bc}{a}\hat{\beta}\end{aligned}$$

■

或者透過估計 $\hat{\alpha}^*$ 與 $\hat{\beta}^*$ 的目標函數, 可得知

$$\min_{\hat{\alpha}^*,\hat{\beta}^*} \sum_{i=1}^{n} (Y_i^* - \hat{\alpha}^* - \hat{\beta}^* X_i^*)^2 \tag{1.1}$$

將 $X_i^* = aX_i + b$, $Y_i^* = cY_i + d$ 代入式 (1.1) 可得到

$$\begin{aligned}\sum_{i=1}^{n} (Y_i^* - \hat{\alpha}^* - \hat{\beta}^* X_i^*)^2 &= \sum_{i=1}^{n} \left[cY_i + d - \hat{\alpha}^* - \hat{\beta}^*(aX_i + b)\right]^2 \\ &= \sum_{i=1}^{n} \left(cY_i + d - \hat{\alpha}^* - b\hat{\beta}^* - \hat{\beta}^* aX_i\right)^2 \quad \xrightarrow{\text{括號內提出 } c} \\ &= c^2 \sum_{i=1}^{n} \left(Y_i - \frac{\hat{\alpha}^* + b\hat{\beta}^* - d}{c} - \frac{\hat{\beta}^* a}{c} X_i\right)^2\end{aligned}$$

令 $\hat{u}_i = Y_i - \hat{\alpha} - \hat{\beta}X_i$, $\hat{u}_i^* = Y_i^* - \hat{\alpha}^* - \hat{\beta}X_i^*$。已知 $\hat{\alpha}, \hat{\beta}$ 可使得 $\sum \hat{u}_i^2$ 有最小值, 因此給定 c 爲一已知常數, 可得知 $\sum \left(\hat{u}_i^*\right)^2$ 之最小值爲 $c^2 \sum \hat{u}_i^2$。其中

$$\sum (\hat{u}_i^*)^2 = \sum_{i=1}^{n} (Y_i^* - \hat{\alpha}^* - \hat{\beta}^* X_i^*)^2 = c^2 \sum_{i=1}^{n} \left(Y_i - \underbrace{\frac{\hat{\alpha}^* + b\hat{\beta}^* - d}{c}}_{\hat{\alpha}} - \underbrace{\frac{\hat{\beta}^* a}{c}}_{\hat{\beta}} X_i\right)^2 = c^2 \sum_{i=1}^{n} \hat{u}_i^2 \tag{1.2}$$

意即若選取之 $\hat{\alpha}^*, \hat{\beta}^*$ 滿足

$$\frac{\hat{\alpha}^* + b\hat{\beta}^* - d}{c} = \hat{\alpha} \tag{1.3}$$

$$\frac{\hat{\beta}^* a}{c} = \hat{\beta} \tag{1.4}$$

則 $\hat{\alpha}^*, \hat{\beta}^*$ 可使得殘差 $\hat{u}_i^*$ 之平方和有最小值。根據上面式 (1.3) 與式 (1.4) 可解出

$$\hat{\beta}^* = \frac{c}{a}\hat{\beta} \tag{1.5}$$

$$\hat{\alpha}^* = c\,\hat{\alpha} + d - \frac{bc}{a}\hat{\beta}$$

■

1. **提供一個簡單的記憶方式**: 因爲 $X_i^* = aX_i + b$, $Y_i^* = cY_i + d$, 令 $\hat{Y}_i = (\hat{Y}_i^* - d)/c$, $X_i = (X_i^* - b)/a$, 帶入樣本迴歸方程式可得到

$$\frac{\hat{Y}_i^* - d}{c} = \hat{\alpha} + \hat{\beta}\frac{X_i^* - b}{a}$$

整理後即可得到

$$\hat{Y}_i^* = \underbrace{c\hat{\alpha} + d - \frac{bc}{a}\hat{\beta}}_{= \hat{\alpha}^*} + \underbrace{\frac{c}{a}\hat{\beta}}_{= \hat{\beta}^*} X_i^*$$

透過改寫迴歸模型, 再比較係數即可得到線性轉換過後的迴歸係數估計值。

2. **特別注意兩個特殊情況:**

   (a) 去除平均 (demean): 即 $(X_i^*, Y_i^*) = (X_i - \bar{X}, Y_i - \bar{Y})$, 則 $\hat{\beta}^* = \hat{\beta}$, $\hat{\alpha}^* = 0$。

   (b) 標準化 (standardize): 即 $(X_i^*, Y_i^*) = \left(\dfrac{X_i - \bar{X}}{S_X}, \dfrac{Y_i - \bar{Y}}{S_Y}\right)$, 則 $\hat{\beta}^* = r_{XY}$, $\hat{\alpha}^* = 0$。

3. 常見的線性轉換有: 單位改變、資料平移 (去除平均) 以及更改虛擬變數的定義等。

4. 線性轉換不會改變相關係數之絕對值, 因此 $r^2_{XY} = r^2_{X^*Y^*}$, 因此**線性轉換不會改變判定係數以及迴歸斜率項係數的顯著性**, 僅會使得斜率項係數與標準誤等比例的變動。將變數進行適當的比例調整使得估計的係數不要太大或太小, 會比較容易解讀其結果。

例題 1.1.1 線性轉換: 解釋變數進行比例變動

(10%) A researcher estimates the following regression equation:

$$Y = 0.5 + 0.006X$$

where Y is the number of asthma hospitalizations, and X is the concentration of sulfur dioxide (SO_2) measured in parts per million (PPM). Suppose the researcher changes the unit of measurement of X from PPM (10^{-6}) to parts per ten thousand (‰), where 1‰ = 10^{-3}, and re-estimates the regression. Answer the following questions:

1. (5%) Write down the new regression equation after the unit change.
2. (5%) Does the coefficient of determination R^2 change? Explain briefly.

《115 台大農經》

《解》

1 已知 $X^* = 0.001X$, 因此

$$\hat{Y} = 0.5 + 0.006X = 0.5 + \frac{0.006}{0.001}X^* = 0.5 + 6X^*$$

2 線性轉換不會改變相關係數的大小, 而判定係數是相關係數的平方, 因此判定係數不會改變。

例題 1.1.2 線性轉換的影響

A change in the unit of measurement of the dependent variable in a model does NOT lead to a change in

(A) The slope coefficient

(B) The standard error of the regression

(C) The standard error of the slope coefficient

(D) The goodness-of-fit of the regression

(E) The confidence interval of the regression 《114 中正財金》

《解》

選 (D)。當變數改變單位, 會造成係數跟標準誤等比例的變動, 但不會改變模型的配適度。

例題 1.1.3 線性轉換: 扣除平均

(8%) Let $E(y) = \mu_Y$, and $E(x) = \mu_X$. Consider two regression models:

$$\text{Model A}: y = \beta_0 + \beta_1 x + \varepsilon,$$
$$\text{Model B}: y^* = \gamma_0 + \gamma_1 x^* + u,$$

where $y^* = y - \mu_y$, $x^* = x - \mu_x$.

Assume $E(\varepsilon) = 0$. What is the relationship between β_1 and γ_1? Explain in detail.

《113 中央財金》

《解》

在原始模型中, $\beta_0 = E(y) - \beta_1 E(x)$, 因此模型可改寫爲

$$y = \mu_y - \beta_1\mu_x + \beta_1 x + \varepsilon$$
$$\underbrace{y - \mu_y}_{y^*} = 0 + \underbrace{\beta_1}_{\gamma_1} \underbrace{(x - \mu_x)}_{x^*} + \varepsilon$$

因此透過比較係數可得知 $\beta_1 = \gamma_1$, 事實上對變數進行平移, 不會改變迴歸的斜率係數。

例題 1.1.4 線性轉換

A sample of 30 weeks (indicated by the subscript t) comes from a small village for us to estimate the following ice cream demand equation:

$$\ln cons_t = \beta_1 + \beta_2 \ln income_t + \beta_3 \ln price_t + \beta_4 tempF_t + \varepsilon_t.$$

The above four time-series variables are logarithmic per-capita weekly consumption of ice cream ($cons_t$, in pints), logarithmic per-capita weekly income ($income_t$, in dollars), logarithmic price of ice cream ($price_t$, in dollars per pint) and average weekly temperature ($tempF_t$, in degrees Fahrenheit), respectively. ε_t is white noise. The OLS estimation result shows:

- $\hat{\beta}_1 = -5.8569$ and $s.e.(\hat{\beta}_1) = 1.3084$
- $\hat{\beta}_2 = 0.7982$ and $s.e.(\hat{\beta}_2) = 0.2605$
- $\hat{\beta}_3 = -0.6262$ and $s.e.(\hat{\beta}_3) = 0.5947$
- $\hat{\beta}_4 = 0.0096$ and $s.e.(\hat{\beta}_4) = 0.0012$
- $R^2 = 0.7405$ and sample size $T = 30$

Where (s.e.(X) denotes the standard error of X). If someone measured temperature in degrees Celsius ($tempC_t$) rather than in degrees Fahrenheit, the ice cream demand equation will be changed to:

$$\ln cons_t = \alpha_1 + \alpha_2 \ln income_t + \alpha_3 \ln price_t + \alpha_4 tempC_t + \varepsilon_t.$$

$$tempC_t = \frac{5}{9} \times (tempF_t - 32)$$

The above alternative estimation of ice cream demand equation will give us $\hat{\alpha}_1$, $\hat{\alpha}_4$ and $s.e.(\hat{\alpha}_4)$ =? 《112 台北經研》

《解》

將 $tempF_t = 9/5 tempC_t + 32$ 代入模型可得到

$$\begin{aligned}\ln cons_t &= -5.8569 + 0.7982 \ln income_t - 0.6262 \ln price_t + 0.0096\left(\frac{9}{5} tempC_t + 32\right)\\ &= -5.5497 + 0.7982 \ln income_t - 0.6262 \ln price_t + 0.01728 tempC_t\end{aligned}$$

因此 $\hat{\alpha}_1 = -5.5497$, $\hat{\alpha}_4 = 0.01728$。而線性轉換不影響 $|t|$ 值的大小, 因此斜率標準誤與斜率估計式等比例的變動, 即

$$SE(\hat{\alpha}_4) = \frac{9}{5} SE(\hat{\beta}_4) = \frac{9}{5} \times 0.0012 = 0.00216$$

例題 1.1.5 僅將解釋變數去除平均

(20%) Assume a linear regression model $Y_i = \beta_0 + \beta_1 X_i^* + \varepsilon_i$, where ε_i are independent $N(0, \sigma^2)$, $i = 1, \ldots..n$, $X_i^* = X_i - \bar{X}$, and $\bar{X} = \sum_{i=1}^{n} X_i / n$

1. (10%) Derive the least squares estimators $(\hat{\beta}_0, \hat{\beta}_1)$ of (β_0, β_1).
2. (10%) Derive the distribution of $\hat{\beta}_0$. 《108 交大管科》

《解》

1 因爲 $X_i^* = X_i - \bar{X}$, 因此

$$\bar{X}^* = \frac{\sum_{i=1}^{n}(X_i - \bar{X})}{n} = 0$$

目標函數爲:

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i^*)^2$$

對目標函數求取一階條件, 可得到

$$\hat{\beta}_0 : \sum_{i=1}^{n} -2(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i^*) = 0$$

$$\hat{\beta}_1 : \sum_{i=1}^{n} -2X_i^*(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i^*) = 0$$

可解出

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i^* - \bar{X}^*)(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i^* - \bar{X}^*)^2} = \frac{\sum_{i=1}^{n} X_i^* (Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i^*)^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}^* = \bar{Y}$$

2 假設 X_i 爲非隨機, 又知 $Y_i = \beta_0 + \beta_1 X_i^* + \varepsilon_i$, 其中 $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, 因此 $Y_i \sim N(\beta_0 + \beta_1 X_i^*, \sigma^2)$ 且相互獨立, 其中 $E(Y_i) = \beta_0 + \beta_1 X_i^*$, $Var(Y_i) = \sigma^2$。又 $\bar{Y}$ 是 Y_i 之線性組合, 因此 $\bar{Y}$ 亦爲常態分配, 其中

$$E(\bar{Y}) = E\left(\frac{\sum_{i=1}^{n} Y_i}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 X_i^*) = \beta_0 + \beta_1 \bar{X}^* = \beta_0$$

$$Var(\bar{Y}) = Var\left(\frac{\sum_{i=1}^{n} Y_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(Y_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

因此可得到

$$\hat{\beta}_0 = \bar{Y} \sim N\left(\beta_0, \frac{\sigma^2}{n}\right)$$

例題 1.1.6 變數進行單位 (比例) 轉換

Please use the following information to answer questions.

A random sample of 100 MBA students is selected from a population and these students' weight and height are recorded. A regression of weight (measured in pounds) on height (measured in inches) yields

$$Weight = -50 + 3.21 \times Height, R^2 = 0.57, SER = 11,$$

where SER refers to the standard error of the regression. Suppose now we change the unit of measurement from pounds to kilograms for weight and from inches to centimeters for height. (Note: 1 pound equals 0.454 kilogram and 1 inch equals 2.54 centimeters.)

1. What is the estimated regression intercept and coefficient of $Height$ in the new regression model?
2. What is the estimated regression SER in the new regression model?
3. What is the estimated regression R^2 in the new regression model?

《109 台大財金甲丙、財金乙》

《解》

[1] 令新變數為 $Weight^*, Height^*$, 已知兩變數的關係為

$$0.454Weight = Weight^*$$
$$2.54Height = Height^*$$

代入迴歸式中可得到

$$\widehat{Weight}^*/0.454 = -50 + 3.21 \times Height^*/2.54$$
$$\widehat{Weight}^* = -22.7 + 0.5737 \times Height^*$$

[2] 由式 1.2 知 $SSE^* = c^2SSE = 0.454^2SSE$, 因此

$$SER^* = \sqrt{\frac{SSE^*}{n-2}} = \sqrt{\frac{0.454^2SSE}{n-2}} = 0.454\underbrace{\sqrt{\frac{SSE}{n-2}}}_{=SER} = 0.454 * 11 = 4.994$$

[3] 線性轉換不會改變迴歸模型的判定係數, 因此 $R^{2*} = R^2 = 0.57$。

例題 1.1.7 線性轉換對於係數的影響

Neneneko recently won the lottery. He plans to spend the money to buy stocks. Neneneko is now studying the property of a certain stock X. Below presents the data of monthly returns of X in the past 8 months (r_x) and the corresponding market returns (r_m) as well as the risk-free rates (r_f). Please answer questions 14 to 16 using the information.

Month_t	1	2	3	4	5	6	7	8
r_x (%)	6	5	1	0	-3	-2	8	13
r_m (%)	3	3	-2	-2	1	1	5	5
r_f (%)	1	1	1	1	1	1	1	1

1. Neneneko is interested in the systematic risk of stock X and would like to estimate the following regression:

$$r_{xt} - r_{ft} = b_0 + b_1(r_{mt} - r_{ft}) + e_t$$

Assume that $e_t \overset{i.i.d.}{\sim} (0, \sigma^2)$ and that $E[e_t | r_{mt}, r_{ft}] = 0$. Which of the followings are correct?

(A) The ordinary least square estimate $\hat{b}_0^{OLS} = 1.533$.

(B) The ordinary least square estimate $\hat{b}_1^{OLS} = 1.672$.

(C) The maximum likelihood estimate $\hat{b}_0^{ML} = 0.014$.

(D) The standard error of $\hat{b}_1^{OLS}$ is 0.513.

2. Ideally, Neneneko should perform the following regression:

$$r_{xt} - r_{ft} = b_0 + b_1(r_{mt} - r_{ft}) + e_t$$

Suppose that Neneneko is careless and mess up his code. He performs the following three regressions instead:

$$r_{xt} = a_0 + a_1(r_{mt} - r_{ft}) + \epsilon_t \quad (1)$$

$$r_{xt} - r_{ft} = c_0 + c_1 r_{mt} + \varepsilon_t \quad (2)$$

$$r_{xt} = d_0 + d_1 r_{mt} + \eta_t \quad (3)$$

Which of the followings are correct?

(A) $\hat{c}_1^{OLS} = \hat{b}_1^{OLS}$

(B) $\hat{d}_0^{OLS} = \hat{b}_0^{OLS} + 1\%$

(C) $\hat{a}_0^{OLS} \neq \hat{b}_0^{OLS} + 1\%$

(D) $\hat{d}_0^{OLS} = \hat{c}_0^{OLS} + 1\%$ 《115 台大財金甲乙丙》

《解》

1 選 (B)、(D)。

Month_t	1	2	3	4	5	6	7	8
$y = r_x - r_f$	0.05	0.04	0	-0.01	-0.04	-0.03	0.07	0.12
$x = r_m - r_f$	0.02	0.02	-0.03	-0.03	0	0	0.04	0.04

根據資料可得到以下統計量

$$\sum X = 0.06 \quad \sum Y = 0.2$$
$$\sum X^2 = 0.0058 \quad \sum Y^2 = 0.026 \quad \sum XY = 0.0097$$

因此可得到

$$\hat{b}_1 = \frac{\sum XY - \sum X \sum Y/n}{\sum X^2 - (\sum X)^2/n} = \frac{0.0097 - 0.06 \times 0.2/8}{0.0058 - (0.06)^2/8} = \frac{164}{107}$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1\bar{X} = 0.025 - \left(\frac{164}{107}\right) \times 0.0075 = 0.0135$$

$$SST = \sum Y^2 - (\sum Y)^2/n = 0.026 - (0.2)^2/8 = 0.021$$

$$SSR = \hat{b}_1^2 \times \left[\sum X^2 - (\sum X)^2/n\right] = \left(\frac{164}{107}\right)^2 (0.0058 - (0.06)^2/8) = 0.0126$$

$$SSE = 0.021 - 0.0126 = 0.0084$$

$$\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{0.0084}{8-2} = 0.0014$$

$$SE(\hat{b}_1) = \sqrt{\frac{\hat{\sigma}^2}{\text{SST}_X}} = \sqrt{\frac{0.0014}{0.0058 - (0.06)^2/8}} = 0.514$$

因為誤差並非來自於常態分配, 因此 ML 估計不一定等於 OLS 估計, (C) 選項錯誤。

2 選 (A)、(D)。平移不影響斜率, 只會影響截距, 因此 $a_1 = b_1 = c_1 = d_1$。簡單的比較係數可得知

$$a_0 = r_{ft} + b_0 = b_0 + 0.01$$

$$c_0 = b_0 - b_1 r_{ft}$$

$$d_0 = b_0 + (1 - b_1)r_{ft} = c_0 + r_{ft} = c_0 + 0.01$$

例題 1.1.8 線性轉換: 當被解釋變數取自然對數時

Consider the following two models:

$$\text{Model I} : \ln Y_i = \beta_1 + \beta_2 X_i + u_i$$
$$\text{Model II} : \ln Y_i^* = \beta_1^* + \beta_2^* X_i^* + u_i$$

where $Y_i^* = w_1 Y_i$, $X_i^* = w_2 X_i$, and w_1 and w_2 are known constants.

1. (5%) Establish the relationships between the coefficient estimators and their standard errors of Models I and II.

2. (5%) Are the R^2 of the two models different? Why or why not? Show your result.

3. (5%) Find the t-statistics for $H_0 : \beta_2 = 0$ (Model I), and for $H_0 : \beta_2^* = 0$ (Model II). Are they the same? Show your result. 《104 中正經研》

《解》

1 將 $Y_i^* = w_1 Y_i$, $X_i^* = w_2 X_i$ 代入迴歸方程式可得到

$$\ln(w_1 Y_i) = \beta_1^* + \beta_2^*(w_2 X_i) + u_i$$

整理過後可得到

$$\ln Y_i = \underbrace{\beta_1^* - \ln w_1}_{\beta_1} + \underbrace{w_2 \beta_2^*}_{\beta_2} X_i + u_i$$

因此

$$\hat{\beta}_1^* = \hat{\beta}_1 + \ln w_1$$
$$\hat{\beta}_2^* = \frac{1}{w_2}\hat{\beta}_2$$

因爲 $\hat{\beta}_2^* = \hat{\beta}_2 / w_2$, 又因兩者應有相同的係數顯著性, 因此

$$SE(\hat{\beta}_2^*) = \frac{1}{|w_2|} SE(\hat{\beta}_2)$$

其中因爲 w_2 可正可負, 故計算標準誤時, w_2 需取絕對值, 以保證 $SE(\hat{\beta}_2^*) > 0$。

2 $\ln Y_i^* = \ln Y_i + \ln w_1$ 是線性轉換, $X_i^* = w_2 X_i$ 亦是線性轉換, 線性轉換不會影響相關係數的大小, 因此轉換過後資料的其相關係數的大小於轉換前相同, 又因 R^2 爲變數之間的相關係數平方, 因此兩個模型的 R^2 會相同。

3 t 檢定統計量的大小只與 R^2 和 n 有關, 因此兩個模型會有相同的顯著性, 惟 t 檢定統計量可能會有正負之分別 (若 $w_2 < 0$)。