**The Ordinary Least Squares (OLS) Estimators**

$$b_2 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} \quad (2.7)$$

$$b_1 = \bar{y} - b_2\bar{x} \quad (2.8)$$

where $\bar{y} = \sum y_i/N$ and $\bar{x} = \sum x_i/N$ are the sample means of the observations on $y$ and $x$.

$$Y = X\beta + e$$

$$\underset{n\times 1}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}} = \underset{n\times n}{\begin{bmatrix} 1 & x_{12} & \cdots & x_{1n} \\ 1 & x_{22} & & \vdots \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & & x_{nn} \end{bmatrix}} \underset{n\times 1}{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}} + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

$$SSE(\beta) = (Y - X\beta)'(Y - X\beta)$$

$$\frac{\partial SSE(\beta)}{\partial \beta} = -2X'(Y - X\beta) = 0$$

$$X'Y - X'X\beta = 0$$

$$X'Y = X'X\beta, \quad \beta = (X'X)^{-1}X'Y$$

$$X'Y = \begin{bmatrix} 1 & \cdots & \cdots & 1 \\ x_{12} & x_{22} & & \\ \vdots & & \ddots & \\ x_{1n} & \cdots & \cdots & x_{nn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$X'X = \begin{bmatrix} 1 & \cdots & \cdots & 1 \\ x_{12} & x_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ x_{1n} & \cdots & \cdots & x_{nn} \end{bmatrix} \begin{bmatrix} 1 & x_{12} & \cdots & x_{1n} \\ \vdots & \ddots & & \\ \vdots & & \ddots & \\ 1 & x_{1n} & \cdots & x_{nn} \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{in} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i2}^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \sum_{i=1}^n x_{in} & \sum_{i=1}^n x_{in}x_{i2} & \cdots & \sum_{i=1}^n x_{in}^2 \end{bmatrix}$$

$$\boxed{K=2}$$

$$X'X = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}$$

$$(X'X)^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}$$

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \frac{1}{\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}$$

$$= \overset{①}{\frac{1}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}} \begin{bmatrix} ③\ \sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i \\ ②\ -\sum x_i \sum y_i + n\sum_{i=1}^n x_i y_i \end{bmatrix}$$

$$= \frac{1}{\frac{1}{n}\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]} \begin{bmatrix} \frac{1}{n}\left[\bar{y}\sum x_i^2 - \bar{x}\sum x_i y_i \right. \\ \frac{1}{n}\left[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right] \end{bmatrix}$$

$$= \begin{bmatrix} \bar{y} - \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}\bar{x} \\ \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} \end{bmatrix} \Rightarrow \begin{cases} b_1 = \bar{y} - b_2\bar{x} \\ b_2 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} \end{cases}$$

<note>

① $\sum(X-\bar{X})^2 = \sum(X^2 - 2X\bar{X} + \bar{X}^2)$

$= \sum X^2 - 2\bar{X}\sum X + n\bar{X}^2$

$= \sum X^2 - n\bar{X}^2 = \frac{1}{n}\left[n\sum X^2 - (\sum X)^2\right]$

② $\sum(X-\bar{X})(Y-\bar{Y})$

$= \sum(XY - \bar{X}Y - X\bar{Y} + \bar{X}\bar{Y})$

$= \sum XY - \bar{X}\sum Y - \bar{Y}\sum X + n\bar{X}\bar{Y}$

$= \sum XY - n\bar{X}\bar{Y} - n\bar{X}\bar{Y} + n\bar{X}\bar{Y}$

$= \sum XY - n\bar{X}\bar{Y}$

$= \sum XY - \frac{1}{n}\sum X \sum Y = \frac{1}{n}(n\sum XY - \sum X \sum Y)$

③ $\bar{Y} - b_2\bar{X} = \bar{Y} - \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}\bar{X}$

$= \bar{Y} - \frac{\bar{X}\sum XY - n\bar{X}^2\bar{Y}}{\sum X^2 - n\bar{X}^2}$

$= \frac{\bar{Y}\sum X^2 - \cancel{n\bar{X}^2\bar{Y}} - \bar{X}\sum XY + \cancel{n\bar{X}^2\bar{Y}}}{\sum X^2 - n\bar{X}^2}$

$= \frac{\bar{Y}\sum X^2 - \bar{X}\sum XY}{\sum X^2 - n\bar{X}^2}$

$$b_1|\mathbf{x} \sim N\left(\beta_1, \frac{\sigma^2 \sum x_i^2}{N\sum(x_i-\bar{x})^2}\right) \tag{2.17}$$

$$b_2|\mathbf{x} \sim N\left(\beta_2, \frac{\sigma^2}{\sum(x_i-\bar{x})^2}\right) \tag{2.18}$$

$$Y \sim N(X\beta, \sigma^2 I)$$

$$\begin{aligned} var(b) &= var[(X'X)^{-1}X'Y] \\ &= (X'X)^{-1}X'\,var(Y)[(X'X)^{-1}X']' \\ &= (X'X)^{-1}X' \cdot \sigma^2 I \cdot X(X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1} \end{aligned}$$

$\boxed{k=2}$

由上題可得 $(X'X)^{-1} = (X'X)^{-1} = \dfrac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix}$

$$= \begin{bmatrix} \dfrac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2} & \dfrac{-\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n}(x_i-\bar{x})^2} \\ \dfrac{-\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n}(x_i-\bar{x})^2} & \dfrac{\cancel{n}}{\cancel{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} \end{bmatrix}$$

$$\begin{array}{ccc} & var(b_1) & cov(b_1, b_2) \\ var(b) = & \left[\begin{array}{c} \dfrac{\sigma^2 \sum x_i^2}{n\sum(x_i-\bar{x})^2} \\ \dfrac{-\bar{x}\sigma^2}{\sum(x_i-\bar{x})^2} \end{array}\right. & \left.\begin{array}{c} \dfrac{-\bar{x}\sigma^2}{\sum(x_i-\bar{x})^2} \\ \dfrac{\sigma^2}{\sum(x_i-\bar{x})^2} \end{array}\right] \\ & & var(b_2) \end{array}$$

**5.3** Consider the following model that relates the percentage of a household's budget spent on alcohol *WALC* to total expenditure *TOTEXP*, age of the household head *AGE*, and the number of children in the household *NK*.

$$WALC = \beta_1 + \beta_2 \ln(TOTEXP) + \beta_3 NK + \beta_4 AGE + e$$

This model was estimated using 1200 observations from London. An incomplete version of this output is provided in Table 5.6.

**TABLE 5.6** Output for Exercise 5.3

Dependent Variable: *WALC*
Included observations: 1200

| Variable | Coefficient | Std. Error | $t$-Statistic | Prob. |
|---|---|---|---|---|
| $C$ | 1.4515 | 2.2019 | 0.6592 | 0.5099 |
| ln(*TOTEXP*) | 2.7648 | 0.4842 | 5.7103 | 0.0000 |
| *NK* | −1.4549 | 0.3695 | −3.9376 | 0.0001 |
| *AGE* | −0.1503 | 0.0235 | −6.4019 | 0.0000 |
| R-squared | 0.0575 | Mean dependent var | | 6.19434 |
| S.E. of regression | 6.2167 | S.D. dependent var | | 6.39547 |
| Sum squared resid | 46221.62 | | | |

**a.** Fill in the following blank spaces that appear in this table.
  i. The $t$-statistic for $b_1$.
  ii. The standard error for $b_2$.
  iii. The estimate $b_3$.
  iv. $R^2$.
  v. $\hat{\sigma}$.

**b.** Interpret each of the estimates $b_2$, $b_3$, and $b_4$.

**c.** Compute a 95% interval estimate for $\beta_4$. What does this interval tell you?

**d.** Are each of the coefficient estimates significant at a 5% level? Why?

**e.** Test the hypothesis that the addition of an extra child decreases the mean budget share of alcohol by 2 percentage points against the alternative that the decrease is not equal to 2 percentage points. Use a 5% significance level.

(a.)

<i> $t = \dfrac{b1}{SE(b1)} = \dfrac{1.4515}{2.2019} = 0.6592$

<ii> $SE(b2) = \dfrac{2.7648}{5.7103} = 0.4842$

<iii> $b3 = 0.3695 \times (-3.9376) = -1.4549$

<v> $SSE = 46221.62$

$MSE = \dfrac{SSE}{n-k} = \dfrac{46221.62}{1200-4} = 38.6468$

$\hat{\sigma} = \sqrt{MSE} = \sqrt{38.6468} = 6.2167$

<iv> $SSTO = S_y^2 \times (n-1) = 6.39547^2 \times 1199$

$R^2 = 1 - \dfrac{SSE}{SSTO} = 1 - \dfrac{46221.62}{6.39547^2 \times 1199} = 0.0575$

---

(b.) b2：家庭 total expenditure 每增加 1%，WALC 增加 2.7648 百分比
b4：家庭 head 每多 1 歲，WALC 下降 0.1503 百分比
b3：家庭 children 每多一位，WALC 下降 1.4549 百分比
(控制其他條件不變下.)

---

(c) C.I. of b4 ⇒ $b4 \pm Z_{0.025} \cdot SE(b4) = [-0.1503 - 1.96 \cdot 0.0235, -0.1503 + 1.96 \cdot 0.0235]$
$= [-0.1964, -0.1042]$

$0 \notin [-0.1964, -0.1042]$，代表年紀對於 WALC，在 95% 信心水準下，有顯著負向影響。

---

(d)
b1 p-value：0.5099 > 0.05 → b1 is insignificant at a 5% level.
b2 p-value：0.0000 < 0.05
b3 p-value：0.0001 < 0.05
b4 p-value：0.0000 < 0.05
} b2, b3, b4 are significant at a 5% level.

---

(e)
$H_0: \beta_3 = -2$
$H_1: \beta_3 \neq -2$ (雙尾)

$\alpha = 0.05$

Test statistic：$t = \dfrac{b3 - (-2)}{SE(b3)} \sim t_{(1196)} \xrightarrow{CLT} Z$

$RR = \{|t| \geq Z_{0.025} = 1.96\}$

$t = \dfrac{-1.4549 + 2}{0.3695} = 1.4752 \notin RR$ ⇒ do not reject H0, the addition of an extra child decrease the mean budget share of alcohol is not significantly differ from 2 percentage.

**5.23** The file *cocaine* contains 56 observations on variables related to sales of cocaine powder in northeastern California over the period 1984–1991. The data are a subset of those used in the study Caulkins, J. P. and R. Padman (1993), "Quantity Discounts and Quality Premia for Illicit Drugs," *Journal of the American Statistical Association*, 88, 748–757. The variables are

*PRICE* = price per gram in dollars for a cocaine sale
*QUANT* = number of grams of cocaine in a given sale
*QUAL* = quality of the cocaine expressed as percentage purity
*TREND* = a time variable with 1984 = 1 up to 1991 = 8
Consider the regression model

$$PRICE = \beta_1 + \beta_2 QUANT + \beta_3 QUAL + \beta_4 TREND + e$$

**a.** What signs would you expect on the coefficients $\beta_2$, $\beta_3$, and $\beta_4$?

**b.** Use your computer software to estimate the equation. Report the results and interpret the coefficient estimates. Have the signs turned out as you expected?

**c.** What proportion of variation in cocaine price is explained jointly by variation in quantity, quality, and time?

**d.** It is claimed that the greater the number of sales, the higher the risk of getting caught. Thus, sellers are willing to accept a lower price if they can make sales in larger quantities. Set up $H_0$ and $H_1$ that would be appropriate to test this hypothesis. Carry out the hypothesis test.

**e.** Test the hypothesis that the quality of cocaine has no influence on expected price against the alternative that a premium is paid for better-quality cocaine.

**f.** What is the average annual change in the cocaine price? Can you suggest why price might be changing in this direction?

(a) I expect the coefficient $\beta_2$ will be negative, the coefficient $\beta_3$ will be positive, and the coefficient $\beta_4$ will be uncertain.

(b) $\widehat{price} = 90.84669 - 0.05997\,QUANT + 0.11621\,QUAL - 2.35458\,TREND$

```
Call:
lm(formula = PRICE ~ QUANT + QUAL + TREND, data = DATA)

Residuals:
    Min      1Q  Median      3Q     Max
-43.479 -12.014  -3.743  13.969  43.753

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept) β1 90.84669    8.58025  10.588 1.39e-14 ***
QUANT       β2 -0.05997    0.01018  -5.892 2.85e-07 ***
QUAL        β3  0.11621    0.20326   0.572   0.5700
TREND       β4 -2.35458    1.38612  -1.699   0.0954 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.06 on 52 degrees of freedom
Multiple R-squared:  0.5097,	Adjusted R-squared:  0.4814
F-statistic: 18.02 on 3 and 52 DF,  p-value: 3.806e-08
```

Holding other variables constant,
for every increase in QUANT, the PRICE drops by 0.05997, as expected.
For every increase in QUAL, the PRICE increases by 0.11621, as expected.
For every increase in TREND, the PRICE drops by $-2.35458$.

(c) $R^2 = \dfrac{SSR}{SST} = 50.97\%$

(d)
$$\begin{cases} H_0: \beta_2 \geq 0 \\ H_1: \beta_2 < 0 \ (\text{左尾}) \end{cases}$$

$\alpha = 0.05$

test statistic: $t = \dfrac{b_2}{SE(b_2)} \sim t(56-4)$

$RR = \{ t \leq -t_{0.05}(52) = -1.675 \}$

$t = \dfrac{-0.05997}{0.01018} = -5.891 \in RR$, reject $H_0$, accept $H_1$, there is a significant evidence to support the claim that the greater number of sales, the lower price they made.

(e)
$$\begin{cases} H_0: \beta_3 = 0 \\ H_1: \beta_3 > 0 \ (\text{右尾}) \end{cases}$$

$\alpha = 0.05$

test statistic: $t = \dfrac{b_3}{SE(b_3)} \sim t(52)$

p-value $= 0.57 > 0.05$, we do not reject $H_0$, there isn't a significant evidence to support the claim that a premium is paid for better-quality cocaine.

(f) Holding other variables constant, every increase in the trend results in a drop of 2.3548, which I think it may be due to the increase in supply, changes in demand, and changes in policy or law enforcement.