The Ordinary Least Squares (OLS) Estimators

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

$$b_1 = \bar{y} - b_2\bar{x} \tag{2.8}$$

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.

Q1: let $K=2$, $Y_i = b_1 + b_2 x_i + \varepsilon_i$, $Y = X\beta + e$

by matrix form

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \beta = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \quad e = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$X'X = \underset{2\times n}{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}} \underset{n\times 2}{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

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

$$X'Y = \underset{2\times n}{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}} \underset{n\times 1}{\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}} = \begin{bmatrix} \sum Y_i \\ \sum x_i Y_i \end{bmatrix}$$

$$b = (X'X)^{-1}(X'Y) = \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \begin{bmatrix} \sum Y_i \\ \sum x_i Y_i \end{bmatrix}$$

$$= \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} \sum x_i^2 \sum Y_i - \sum x_i \sum x_i Y_i \\ -\sum x_i \sum Y_i + n\sum x_i Y_i \end{bmatrix}$$

$$\therefore b_1 = \frac{\Sigma x_i^2 \Sigma y_i - \Sigma x_i \Sigma x_i y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2} \qquad b_2 = \frac{-\Sigma x_i \Sigma y_i + n\Sigma x_i y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2}$$

by normal form

$$b_2 = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2} = \frac{\Sigma(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y})}{\Sigma x_i^2 - n\bar{x}^2}$$

$$= \frac{\Sigma x_i y_i - \bar{y}\Sigma x_i - \bar{x}\Sigma y_i + n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}^2}$$

$$= \frac{\Sigma x_i y_i - n\bar{y}\bar{x} \cancel{- n\bar{x}\bar{y}} \cancel{+ n\bar{x}\bar{y}}}{\Sigma x_i^2 - n\bar{x}^2}$$

$$= \frac{\Sigma x_i y_i - n\left(\frac{\Sigma y_i}{n}\right)\left(\frac{\Sigma x_i}{n}\right)}{\Sigma x_i^2 - n\left(\frac{\Sigma x_i}{n}\right)^2} \quad \text{同乘 } n$$

$$= \frac{n\Sigma x_i y_i - \Sigma y_i \Sigma x_i}{n\Sigma x_i^2 - (\Sigma x_i)^2} \quad \#$$

$$b_1 = \bar{y} - b_2\bar{x} = \frac{\bar{y}\Sigma(x_i-\bar{x})^2 - \bar{x}\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2}$$

$$= \frac{\bar{y}\sum x_i^2 - \cancel{n\bar{y}\bar{x}^2} - \bar{x}\sum x_i y_i + \cancel{\bar{x}n\bar{y}\bar{x}}}{\sum x_i^2 - n\bar{x}^2}$$

$$= \frac{\frac{\sum y_i}{n}\sum x_i^2 - \frac{\sum x_i}{n}\sum x_i y_i}{\sum x_i^2 - n\left(\frac{\sum x_i}{n}\right)^2} \text{ 同乘 } n$$

$$= \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \quad \#$$

Q2:

$$\operatorname{var}(b_1|\mathbf{x}) = \sigma^2\left[\frac{\sum x_i^2}{N\sum(x_i-\bar{x})^2}\right] \tag{2.14}$$

$$\operatorname{var}(b_2|\mathbf{x}) = \frac{\sigma^2}{\sum(x_i-\bar{x})^2} \tag{2.15}$$

$$\operatorname{cov}(b_1,b_2|\mathbf{x}) = \sigma^2\left[\frac{-\bar{x}}{\sum(x_i-\bar{x})^2}\right] \tag{2.16}$$

$$\operatorname{Var}(b) = \sigma^2(X'X)^{-1} = \sigma^2 \frac{1}{n\sum x_i^2 - (\sum x_i)^2}\begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\sigma^2\sum x_i^2}{n\sum x_i^2 - (\sum x_i)^2} & \frac{-\sigma^2\sum x_i}{n\sum x_i^2 - (\sum x_i)^2} \\ \frac{-\sigma^2\sum x_i}{n\sum x_i^2 - (\sum x_i)^2} & \frac{\sigma^2 n}{n\sum x_i^2 - (\sum x_i)^2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\sigma^2\sum x_i^2}{n\sum(x_i-\bar{x})^2} & \frac{-\sigma^2\bar{x}}{\sum(x_i-\bar{x})^2} \\ \frac{-\sigma^2\bar{x}}{\sum(x_i-\bar{x})^2} & \frac{\sigma^2}{\sum(x_i-\bar{x})^2} \end{bmatrix}$$

$$= \begin{bmatrix} \operatorname{Var}(b_1|X) & \operatorname{cov}(b_1,b_2|X) \\ \operatorname{cov}(b_1,b_2|X) & \operatorname{Var}(b_2|X) \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{bmatrix}$$

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.

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.** $\hat{\sigma}$.

(a)

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	6.217	S.D. dependent var (S_Y)		6.39547
Sum squared resid	46221.62			

t-statistic for $b_1 = 1.4515 \div 2.2019 = 0.6592$

standard error for $b_2 = 2.7648 \div 5.7103 = 0.4842$

the estimate $b_3 = 0.3695 \times (-3.9376) = -1.4549$

$$SST = (n-1)S_{YY} = (n-1)S_Y^2 = 1199 \cdot (6.39547)^2 = 49041.5418$$

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{46221.62}{49041.5418} = 0.0575$$

$$\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{n-K}} = \sqrt{\frac{46221.62}{1196}} \approx 6.217$$

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

$b_2 = 2.765$, other conditions hold, 1% increase in total expenditure will increase the expenditure going to alcohol WALC by 2.765%

$b_3 = -1.4549$ other conditions hold, 1 more child the household has, will decrease the expenditure going to alcohol WALC by 1.4549 units

$b_4 = -0.1503$ other conditions hold, 1 year increase of the household age, will decrease the expenditure going to alcohol WALC by 0.1503 units

c. Compute a 95% interval estimate for β_4. What does this interval tell you?

95% Interval estimate =

$[\hat{\beta}_4 - t_{0.025,1196} \cdot SE(\hat{\beta}_4), \hat{\beta}_4 + t_{0.025,1196} \cdot SE(\hat{\beta}_4)]$

$= [-0.1503 - 1.96 \cdot 0.0235, -0.1503 + 1.96 \cdot 0.0235]$

$= [-0.1964, -0.1042]$

我們有95%的信心 β_4 會落在 $[-0.1964, -0.1042]$ 之間

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

Except for intercept, all coefficient estimates are significantly different from 0 at 5% level ($\because$ p-value < 0.05)

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.

$H_0: \beta_3 = -2$, $H_1: \beta_3 \neq -2$

$$T_0 = \frac{-1.4549-(-2)}{0.3695} \approx 1.4752$$

$$RR = \{T_0 < t_{0.025,1196} \text{ or } T_0 > t_{0.975,1196}\}$$
$$\approx \{T_0 < -1.96 \text{ or } T_0 > 1.96\}$$

$\because T_0 \notin RR$ $\therefore$ don't reject H_0

There is no evidence to say the decrease is not equal to 2

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 β_2, β_3, and β_4?

(a) We expect β_2 is negative because the number of grams increase. the price per gram should decrease

β_3 is positive because the purer the cocaine, the higher the price

β_4 is uncertain depend on how demand and supply are changing over time

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?

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

```
Call:
lm(formula = price ~ quant + qual + trend, data = cocaine)

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) 90.84669    8.58025  10.588 1.39e-14 ***
quant       -0.05997    0.01018  -5.892 2.85e-07 ***
qual         0.11621    0.20326   0.572   0.5700
trend       -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
```

the other conditions hold, increase 1 unit of quantity, the price will decrease 0.05997 unit

the other conditions hold, increase 1 unit of quality, the price will increase 0.11621 unit

the other conditions hold, increase 1 year of time, the price will decrease 2.3546 unit

All the signs turn out to be the same to our expectation. β4 implies that supply has been increasing faster than demand

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

(C) $R^2 = 0.5097$ (模型可解釋部分)

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.

(d) $H_0: \beta_2 \geq 0 \quad H_1: \beta_2 < 0 \quad , \alpha = 5\%$

$$T_0: \frac{-0.05997 - 0}{0.01018} = -5.89$$

$$RR = \{ T_0 < -t_{(0.95, 56-4)} = -1.675 \}$$

$\because T_0 \in RR$, we reject H_0, the sellers are willing to accept a lower price if they can make sales in larger quantities.

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.

(e) $H_0: \beta_3 \leq 0 \ , H_1: \beta_3 > 0 \quad \alpha = 5\%$

$$T_0: \frac{0.11621 - 0}{0.20326} = 0.5717$$

$$RR = \{ T_0 > t_{(0.95, 52)} = 1.675 \}$$

$\because T_0 \notin RR$, $\therefore$ there is no evidence to say 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?

(f) average annual change depend on b4 = -2.35458, the price is decreasing over time, the reason might be the develope technology of producing cocaine, but the demain remains the same, leads to supply exceeds demand so the price fall.