2.11 Exercises

2.11.1 Problems

2.1 Consider the following five observations. You are to do all the parts of this exercise using only a calculator.

x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
3	4	2	4	2	4
2	2	1	1	0	0
1	3	0	0	1	0
−1	1	−2	4	−1	2
0	0	−1	1	−2	2
$\sum x_i =$ 5	$\sum y_i =$ 10	$\sum(x_i-\bar{x}) =$ 0	$\sum(x_i-\bar{x})^2 =$ 10	$\sum(y_i-\bar{y}) =$ 0	$\sum(x_i-\bar{x})(y_i-\bar{y}) =$ 8

a. Complete the entries in the table. Put the sums in the last row. What are the sample means $\bar{x}$ and $\bar{y}$?

b. Calculate b_1 and b_2 using (2.7) and (2.8) and state their interpretation.

c. Compute $\sum_{i=1}^{5} x_i^2, \sum_{i=1}^{5} x_i y_i$. Using these numerical values, show that $\sum(x_i-\bar{x})^2 = \sum x_i^2 - N\bar{x}^2$ and $\sum(x_i-\bar{x})(y_i-\bar{y}) = \sum x_i y_i - N\bar{x}\bar{y}$.

d. Use the least squares estimates from part (b) to compute the fitted values of y, and complete the remainder of the table below. Put the sums in the last row.
Calculate the sample variance of y, $s_y^2 = \sum_{i=1}^{N}(y_i-\bar{y})^2/(N-1)$, the sample variance of x, $s_x^2 = \sum_{i=1}^{N}(x_i-\bar{x})^2/(N-1)$, the sample covariance between x and y, $s_{xy} = \sum_{i=1}^{N}(y_i-\bar{y})(x_i-\bar{x})/(N-1)$, the sample correlation between x and y, $r_{xy} = s_{xy}/(s_x s_y)$ and the coefficient of variation of x, $CV_x = 100(s_x/\bar{x})$. What is the median, 50th percentile, of x?

x_i	y_i	$\hat{y}_i$	$\hat{e}_i$	$\hat{e}_i^2$	$x_i\hat{e}_i$
3	4				
2	2				
1	3				
−1	1				
0	0				
$\sum x_i =$	$\sum y_i =$	$\sum \hat{y}_i =$	$\sum \hat{e}_i =$	$\sum \hat{e}_i^2 =$	$\sum x_i\hat{e}_i =$

e. On graph paper, plot the data points and sketch the fitted regression line $\hat{y}_i = b_1 + b_2 x_i$.

f. On the sketch in part (e), locate the point of the means $(\bar{x}, \bar{y})$. Does your fitted line pass through that point? If not, go back to the drawing board, literally.

g. Show that for these numerical values $\bar{y} = b_1 + b_2\bar{x}$.

h. Show that for these numerical values $\hat{\bar{y}} = \bar{y}$, where $\hat{\bar{y}} = \sum \hat{y}_i/N$.

i. Compute $\hat{\sigma}^2$.

j. Compute $\widehat{\text{var}}(b_2|\mathbf{x})$ and $\text{se}(b_2)$.

2.2 A household has weekly income of \$2000. The mean weekly expenditure for households with this income is $E(y|x = \$2000) = \mu_{y|x=\$2000} = \$220$, and expenditures exhibit variance $\text{var}(y|x = \$2{,}000) = \sigma^2_{y|x=\$2{,}000} = \$121$.

a. Assuming that weekly food expenditures are normally distributed, find the probability that a household with this income spends between \$200 and \$215 on food in a week. Include a sketch with your solution.

f. Using the data in Table 2.4, calculate the sum of squared residuals $S\left(\hat{\beta}_1\right) = \sum_{i=1}^{N}\left(y_i - \hat{\beta}_1\right)^2$. Is this sum of squared residuals larger or smaller than the sum of squared residuals $S\left(b_1, b_2\right) = \sum_{i=1}^{N}\left(y_i - b_1 - b_2 x_i\right)^2$ using the least squares estimates? [See Exercise 2.3 (d).]

2.11 Let y = expenditure ($) on food away from home per household member per month in the past quarter and x = monthly household income (in hundreds of dollars) during the past year.

a. Using 2013 data from three-person households ($N = 2334$), we obtain least squares estimates $\hat{y} = 13.77 + 0.52x$. Interpret the estimated slope and intercept from this relation.

b. Predict the expenditures on food away from home for a household with $2000 a month income.

c. Calculate the elasticity of expenditure on food away from home with respect to income when household income is $2000 per month. [*Hint*: Elasticity must be calculated for a point on the fitted regression.]

d. We estimate the log-linear model to be $\widehat{\ln(y)} = 3.14 + 0.007x$. What is the estimated elasticity of expenditure on food away from home with respect to income, if household income is $2000 per month?

e. For the log-linear model in part (d), calculate $\hat{y} = \exp(3.14 + 0.007x)$ when $x = 20$ and when $x = 30$. Evaluate the slope of the relation between y and x, dy/dx, for each of these $\hat{y}$ values. Based on these calculations for the log-linear model, is expenditure on food away from home increasing with respect to income at an increasing or decreasing rate?

f. When estimating the log-linear model in part (d), the number of observations used in the regression falls to $N = 2005$. How many households in the sample reported no expenditures on food away from home in the past quarter?

2.12 Let y = expenditure ($) on food away from home per household member per month in the past quarter and $x = 1$ if the household includes a member with an advanced degree, a Master's, or Ph.D./Professional degree, and $x = 0$ otherwise.

a. Using 2013 data from three-person households ($N = 2334$), we obtain least squares estimates $\hat{y} = 44.96 + 30.41x$. Interpret the coefficient of x and the intercept from this relation.

b. What is the per person sample mean of food expenditures away from home for a household including someone with an advanced degree?

c. What is the per person sample mean of food expenditures away from home for a household that does not include someone with an advanced degree?

2.13 Using 2011 data on 141 U.S. public research universities, we examine the relationship between academic cost per student, *ACA* (real total academic cost per student in thousands of dollars) and full-time enrollment *FTESTU* (in thousands of students).

a. The least squares fitted relation is $\widehat{ACA} = 14.656 + 0.266FTESTU$. What is the economic interpretation of the estimated parameters? Why isn't the intercept zero?

b. In 2011 Louisiana State University (LSU) had a full-time student enrollment of 27,950. Using the fitted related in part (a), compute the predicted value of *ACA*.

c. The actual value of *ACA* for LSU that year was 21.403. Calculate the least squares residual for LSU? Does the model overpredict or underpredict *ACA* for LSU?

d. The sample mean (average) full-time enrollment in U.S. public research universities in 2011 was 22,845.77. What was the sample mean of academic cost per student?

2.14 Consider the regression model $WAGE = \beta_1 + \beta_2 EDUC + e$, where *WAGE* is hourly wage rate in U.S. 2013 dollars and *EDUC* is years of education, or schooling. The regression model is estimated twice using the least squares estimator, once using individuals from an urban area, and again for individuals in a rural area.

$$\text{Urban} \quad \underset{(\text{se})}{\widehat{WAGE} = -10.76 + 2.46\ EDUC,} \quad \underset{(2.27)\ \ (0.16)}{N = 986}$$

$$\text{Rural} \quad \underset{(\text{se})}{\widehat{WAGE} = -4.88 + 1.80\ EDUC,} \quad \underset{(3.29)\ \ (0.24)}{N = 214}$$

a. Using the estimated rural regression, compute the elasticity of wages with respect to education at the "point of the means." The sample mean of *WAGE* is $19.74.

b. The sample mean of *EDUC* in the urban area is 13.68 years. Using the estimated urban regression, compute the standard error of the elasticity of wages with respect to education at the "point of the means." Assume that the mean values are "givens" and not random.

c. What is the predicted wage for an individual with 12 years of education in each area? With 16 years of education?

2.15 Professor E.Z. Stuff has decided that the least squares estimator is too much trouble. Noting that two points determine a line, Dr. Stuff chooses two points from a sample of size N and draws a line between them, calling the slope of this line the EZ estimator of β_2 in the simple regression model. Algebraically, if the two points are (x_1, y_1) and (x_2, y_2), the EZ estimation rule is

$$b_{EZ} = \frac{y_2 - y_1}{x_2 - x_1}$$

Assuming that all the assumptions of the simple regression model hold:

a. Show that b_{EZ} is a "linear" estimator.

b. Show that b_{EZ} is an unbiased estimator.

c. Find the conditional variance of b_{EZ}.

d. Find the conditional probability distribution of b_{EZ}.

e. Convince Professor Stuff that the EZ estimator is not as good as the least squares estimator. No proof is required here.

2.11.2 Computer Exercises

2.16 The capital asset pricing model (CAPM) is an important model in the field of finance. It explains variations in the rate of return on a security as a function of the rate of return on a portfolio consisting of all publicly traded stocks, which is called the *market* portfolio. Generally, the rate of return on any investment is measured relative to its opportunity cost, which is the return on a risk-free asset. The resulting difference is called the *risk premium*, since it is the reward or punishment for making a risky investment. The CAPM says that the risk premium on security j is *proportional* to the risk premium on the market portfolio. That is,

$$r_j - r_f = \beta_j(r_m - r_f)$$

where r_j and r_f are the returns to security j and the risk-free rate, respectively, r_m is the return on the market portfolio, and β_j is the jth security's "*beta*" value. A stock's *beta* is important to investors since it reveals the stock's volatility. It measures the sensitivity of security j's return to variation in the whole stock market. As such, values of *beta* less than one indicate that the stock is "defensive" since its variation is less than the market's. A *beta* greater than one indicates an "aggressive stock." Investors usually want an estimate of a stock's *beta* before purchasing it. The CAPM model shown above is the "economic model" in this case. The "econometric model" is obtained by including an intercept in the model (even though theory says it should be zero) and an error term

$$r_j - r_f = \alpha_j + \beta_j(r_m - r_f) + e_j$$

a. Explain why the econometric model above is a simple regression model like those discussed in this chapter.

b. In the data file *capm5* are data on the monthly returns of six firms (GE, IBM, Ford, Microsoft, Disney, and Exxon-Mobil), the rate of return on the market portfolio (*MKT*), and the rate of return on the risk-free asset (*RISKFREE*). The 180 observations cover January 1998 to December 2012. Estimate the CAPM model for each firm, and comment on their estimated *beta* values. Which firm appears most aggressive? Which firm appears most defensive?

c. Finance theory says that the intercept parameter α_j should be zero. Does this seem correct given your estimates? For the Microsoft stock, plot the fitted regression line along with the data scatter.

d. Estimate the model for each firm under the assumption that $\alpha_j = 0$. Do the estimates of the *beta* values change much?

2.17 The data file *collegetown* contains observations on 500 single-family houses sold in Baton Rouge, Louisiana, during 2009–2013. The data include sale price (in thousands of dollars), *PRICE*, and total interior area of the house in hundreds of square feet, *SQFT*.

a. Plot house price against house size in a scatter diagram.

2.1

x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
3	4	2	4	2	4
2	2	1	1	0	0
1	3	0	0	1	0
−1	1	−2	4	−1	2
0	0	−1	1	−2	2
$\sum x_i = 5$	$\sum y_i = 10$	$\sum(x_i-\bar{x}) = 0$	$\sum(x_i-\bar{x})^2 = 10$	$\sum(y_i-\bar{y}) = 0$	$\sum(x_i-\bar{x})(y_i-\bar{y}) = 8$
1	2				

a. $\bar{x} = \frac{3+2+1+(-1)+0}{5} = 1$

$\bar{y} = \frac{4+2+3+1+0}{5} = 2$

b. b_1：截距項，代表 $x=0$ 時，y 之分

b_2：斜率，代表 x 增加 1 單位，y 增加 b_2 分

$b_2 = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = 0.8$

$b_1 = \bar{Y} - b_2\bar{X} = 2 - 0.8(1) = 1.2$

c. $\sum_{i=1}^{5} x_i^2 = 9+4+1+1 = 15$

$\sum_{i=1}^{5} x_i y_i = 12+4+3-1 = 18$

$\Sigma(x_i-\bar{x})^2 = 10$

$\Sigma x_i^2 - n\bar{X} = 15 - 5(1) = 10$，故得證

$\Sigma(x_i-\bar{x})(y_i-\bar{y}) = 8$

$\Sigma x_i y_i - n\bar{x}\bar{y} = 18 - 5(2) = 8$，故得證

d. $S_y^2 = \frac{\Sigma(y_i-\bar{y})^2}{N-1} = \frac{2^2+0^2+1^2+(-1)^2+(-2)^2}{4} = 2.5$

$S_x^2 = \frac{\Sigma(x_i-\bar{x})^2}{N-1} = \frac{2^2+1^2+0+(-2)^2+(-1)^2}{4} = 2.5$

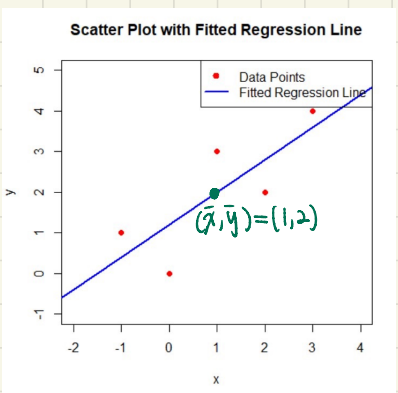
$S_{xy} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{N-1} = \frac{8}{4} = 2$

$r_{xy} = \frac{S_{xy}}{S_x S_y} = 0.8$

$CV_x = 100(S_x/\bar{x}) = 100\left(\frac{\sqrt{2.5}}{1}\right) \approx 158.11$

median $(x) = 1$

e.



f. Yes, the fitted line pass through $(\bar{x},\bar{y}) = (1,2)$

g. $\bar{\hat{y}} = b_1 + b_2\bar{X} = 1.2 + 0.8(1) = 2$

h. $\bar{\hat{y}} = \frac{\Sigma\hat{y}_i}{N} = \frac{4+2+3+1+0}{5} = 2 = \bar{y}$

i. $\hat{\sigma}^2 = \frac{\sum_{i=1}^{5}(y_i-\hat{y}_i)^2}{N-2} = \frac{3.6}{3} = 1.2$

j. $Var(b_2|x) = \hat{\sigma}^2 \cdot \frac{1}{\Sigma(x_i-\bar{x})^2} = 1.2 \times \frac{1}{10} = 0.12$

$\hat{y} = 1.2 + 0.8x$

$\hat{y}_i = (0.4, 1.2, 2, 2.8, 3.6)$

$y_i = (1, 0, 3, 2, 4)$

$-0.6 \quad 1.2 \quad | \quad 0.8 \quad 0.4$

$0.36 \quad 1.44 \quad | \quad 0.64 \quad 0.16$

$$se(b_2) = \sqrt{Var(b_2|X)} = \sqrt{0.12} \approx 0.3464$$

2.14

a.

$$\overline{WAGE}_R = 19.74$$

$$19.74 = -4.88 + 1.80\,\overline{EDUC}_R$$

$$\overline{EDUC}_R = 13.6778$$

$$\hat{\varepsilon}_R = \frac{\partial y}{\partial x} \cdot \frac{x}{y} = b_1 \times \frac{\overline{EDUC}_R}{\overline{WAGE}_R} = 1.8 \times \frac{13.6778}{19.74} = 1.2472$$

b.

$$\overline{EDUC}_U = 13.68$$

$$\overline{WAGE}_U = -10.76 + 2.46 \times 13.68 = 22.8928$$

$$SE(b_2) = 0.16$$

$$\text{标准误} = SE(\hat{\varepsilon}_U) = SE\left(b_2 \times \frac{\overline{EDUC}_U}{\overline{WAGE}_U}\right) = \frac{\overline{EDUC}_U}{\overline{WAGE}_U} \times SE(b_2) = \frac{13.68}{22.8928} \times 0.16 = 0.0956$$

c.

	12 Years:	16 Years:
Urban:	$-10.76+2.46(12)$ $= 18.76$	$-10.76+2.46(16)$ $= 28.6$
Rural:	$-4.88+1.80(12)$ $= 16.72$	$-4.88+1.8(16)$ $= 23.92$

2.16

a.

$$r_i - r_f = \alpha + \beta(r_m - r_f) + e_j$$

對應 ↓ ↓ ↓ ↓ ↓

y b_1 b_2 x ε_i

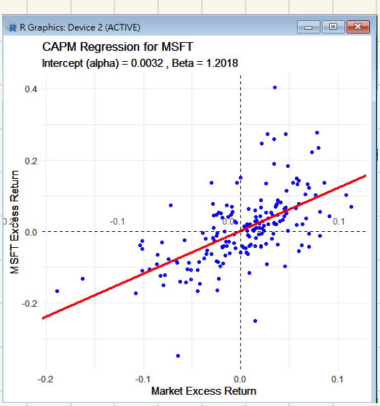
故符合簡迴歸 $y = b_1 + b_2 x + \varepsilon$ 形式

b.

由圖可知，Ford 是最 aggressive 的，∵ $\beta = 1.66$ 為最大；

Exxon 則是最 defensive 的，∵ $\beta = 0.46$ 為最小。

c.



從圖中觀察，可發現 MSFT 有截距 0.0032

但從統計數據中可發現截距項不顯著。

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.003250   0.006036   0.538    0.591
Market_Excess 1.201840   0.122152   9.839   <2e-16 ***
---
```

d.

	Company	Beta_With_Intercept	Beta_No_Intercept
1	ge	1.1479521	1.1467633
2	ibm	0.9768898	0.9843954
3	ford	1.6620307	1.6667168
4	msft	1.2018398	1.2058695
5	dis	1.0115207	1.0128190
6	xom	0.4565208	0.4630727

可以觀察到有無考慮 interception 對 β 的估計影響不大，Rm-Rf 已經是很好的估計 β 的變數了。