3.1 There were 64 countries in 1992 that competed in the Olympics and won at least one medal. Let *MEDALS* be the total number of medals won, and let *GDPB* be GDP (billions of 1995 dollars). A linear regression model explaining the number of medals won is $MEDALS = \beta_1 + \beta_2 GDPB + e$. The estimated relationship is

$$\widehat{MEDALS} = b_1 + b_2 GDPB = \underset{(2.38994)}{7.61733} + \underset{(0.00215)}{0.01309} GDPB \qquad \text{(se)} \qquad \text{(XR3.1)}$$

a. We wish to test the hypothesis that there is no relationship between the number of medals won and *GDP* against the alternative there is a positive relationship. State the null and alternative hypotheses in terms of the model parameters.

b. What is the test statistic for part (a) and what is its distribution if the null hypothesis is true?

c. What happens to the distribution of the test statistic for part (a) if the alternative hypothesis is true? Is the distribution shifted to the left or right, relative to the usual t-distribution? [*Hint*: What is the expected value of b_2 if the null hypothesis is true, and what is it if the alternative is true?]

d. For a test at the 1% level of significance, for what values of the t-statistic will we reject the null hypothesis in part (a)? For what values will we fail to reject the null hypothesis?

e. Carry out the t-test for the null hypothesis in part (a) at the 1% level of significance. What is your economic conclusion? What does 1% level of significance mean in this example?

(a)

虛無假設 H0: b2 = 0，表示 GDP 與奧運獎牌數沒有關係

對立假設 H1: b2 > 0，表示 GDP 越高，獲得的奧運獎牌數越多

(b)

t-statistic = (b2 – 0) / SE = (0.01309 – 0) / 0.00215 = 6.09

在自由度 = n-2 的 t 分布下檢定，n = 64，所以自由度等於 62。

如果 H0 為真，t-統計量服從 t(62)的分布。

(c)

如果 H1 為真，則 t-統計量的分佈會向右偏移。

(d)

$t_{0.99, 62} = 2.39$ (查表)

拒絕域為 $t > 2.39$，如果 t 值大於 2.39，就拒絕 H0，反之則不拒絕 H0

(e)

由(b)得出的 t 值為 6.09，落在拒絕域中，表示 GDP 確實對獎牌數有顯著的正向影響。

表示 GDP 較高的國家通常會贏得更多獎牌，GDPB 每增加一單位，獎牌會增加 0.13 面。

1% 顯著水準表示我們有 99%的信心 這個結果不是隨機誤差造成的。在 H0 成立的情況下，發生這種極端結果的機率只有 1%

3.7 We have 2008 data on *INCOME* = income per capita (in thousands of dollars) and *BACHELOR* = percentage of the population with a bachelor's degree or more for the 50 U.S. States plus the District of Columbia, a total of $N = 51$ observations. The results from a simple linear regression of *INCOME* on *BACHELOR* are

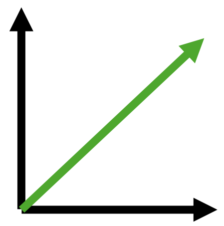
$$\begin{array}{llll} \widehat{INCOME} = & (a) & + & 1.029 BACHELOR \\ \text{se} & (2.672) & & (c) \\ t & (4.31) & & (10.75) \end{array}$$

a. Using the information provided calculate the estimated intercept. Show your work.
b. Sketch the estimated relationship. Is it increasing or decreasing? Is it a positive or inverse relationship? Is it increasing or decreasing at a constant rate or is it increasing or decreasing at an increasing rate?
c. Using the information provided calculate the standard error of the slope coefficient. Show your work.
d. What is the value of the t-statistic for the null hypothesis that the intercept parameter equals 10?
e. The p-value for a two-tail test that the intercept parameter equals 10, from part (d), is 0.572. Show the p-value in a sketch. On the sketch, show the rejection region if $\alpha = 0.05$.
f. Construct a 99% interval estimate of the slope. Interpret the interval estimate.
g. Test the null hypothesis that the slope coefficient is one against the alternative that it is not one at the 5% level of significance. State the economic result of the test, in the context of this problem.

(a)

a = t × SE = 4.31 × 2.672 = 11.52

(b)



為正向關係，因為 BACHELOR 的係數為正，並以固定速率上升

(c)

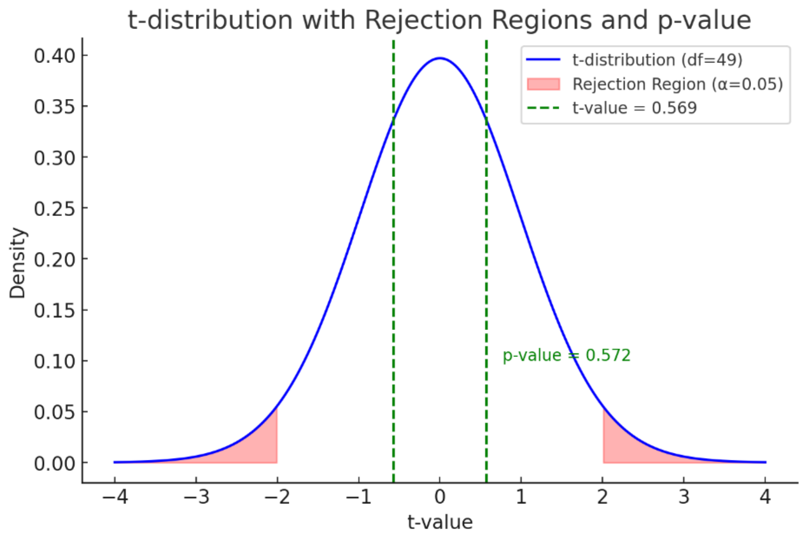
SE = b2 / t = 1.029 / 10.75 = 0.0958。

斜率的標準誤為 0.0958

(d)

H0: a = 10，H1: a != 10。t = (11.52 – 10) / 2.672 = 0.569

(e)



在 α = 0.05 時，拒絕區間在兩側 t 值超過 1.96 或 -1.96 的地方

而 p 值 0.572 太大，不會落在拒絕區間，所以我們無法拒絕零假設

(f)

自由度 = 49。$t_{0.005, 49}$ = 2.68。1.029 ± (2.68 × 0.0958) = [0.772, 1.286]

代表在 99% 的信心水準之下，每多 1%人口擁有學士學位，收入的影響範圍在

0.772 到 1.286 千美元之間

(g)

H0: b2 = 1，H1: b2 != 1。t = 0.029 / 0.0958 = 0.303

$\alpha = 0.05$ 時，對應的 t 分布雙尾臨界值為 1.96，因為 t 值小於 1.96，無法拒絕 H_0

代表學歷每增加 1%，收入增加約 1.029 千美元，跟學歷每增加 1%，收入增加 1 千美元，兩者之間沒有統計上的顯著差異

教育對收入的影響約為 1：1，教育每增加 1%，收入也增加 1%

3.17 Consider the regression model $WAGE = \beta_1 + \beta_2 EDUC + e$. Where $WAGE$ is hourly wage rate in US 2013 dollars. $EDUC$ is years of schooling. The model is estimated twice, once using individuals from an urban area, and again for individuals in a rural area.

$$\text{Urban} \qquad \begin{array}{l} \widehat{WAGE} = -10.76 + 2.46 EDUC, \quad N = 986 \\ \text{(se)} \qquad\quad (2.27) \quad (0.16) \end{array}$$

$$\text{Rural} \qquad \begin{array}{l} \widehat{WAGE} = -4.88 + 1.80 EDUC, \quad N = 214 \\ \text{(se)} \qquad\quad (3.29) \quad (0.24) \end{array}$$

a. Using the urban regression, test the null hypothesis that the regression slope equals 1.80 against the alternative that it is greater than 1.80. Use the $\alpha = 0.05$ level of significance. Show all steps, including a graph of the critical region and state your conclusion.

b. Using the rural regression, compute a 95% interval estimate for expected $WAGE$ if $EDUC = 16$. The required standard error is 0.833. Show how it is calculated using the fact that the estimated covariance between the intercept and slope coefficients is -0.761.

c. Using the urban regression, compute a 95% interval estimate for expected $WAGE$ if $EDUC = 16$. The estimated covariance between the intercept and slope coefficients is -0.345. Is the interval estimate for the urban regression wider or narrower than that for the rural regression in (b). Do you find this plausible? Explain.

d. Using the rural regression, test the hypothesis that the intercept parameter β_1 equals four, or more, against the alternative that it is less than four, at the 1% level of significance.

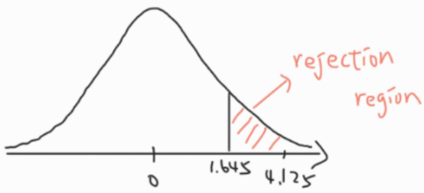
(a)

$H_0: b_2 = 1.80$, $H_1: b_2 > 1.80$, $\alpha = 0.05$

$t = \dfrac{2.46 - 1.8}{0.16} \approx 4.125 > 1.645 = t_{0.05, 984}$

所以拒絕 H_0

表示當EDUC增加一單位，WAGE會增加超過1.8



(b) $EDUC = 16$, $\widehat{WAGE} = -4.88 + 1.8 \times 16 = 23.92$

$t_{0.025, 212} = 1.97$

$CI = 23.92 \pm 1.97 \times \underline{0.833} = [22.28, 25.56]$

$SE(WAGE) = \sqrt{3.29^2 + 16^2 \times 0.24^2 + 2 \times 16 \times (-0.761)}$

(c) $EDUC = 16$, $\widehat{WAGE} = -10.76 + 2.46 \times 16 = 28.6$

$SE(WAGE) = \sqrt{2.27^2 + 16^2 \times 0.16^2 + 2 \times 16 \times (-0.345)} = 0.816$

$t_{0.025, 984} = 1.96$

$CI = 28.6 \pm 1.96 \times 0.816 = [26.99, 30.21]$

Rural's CI width = 3.28

Urban's CI width = 3.22

Urban的寬度較窄，因為樣本較大

估計的不確定性較小

(d)

$H_0: \beta_1 \geq 4$, $H_1: \beta_1 < 4$. $\alpha = 0.01$

$$t = \frac{\hat{\beta}_1 - 4}{se(\hat{\beta}_1)} = \frac{(-4.88-4)}{3.29} = -2.7$$

$t_{0.01,212} = -2.33$, $t < -2.33$. reject H_0. 統計結果顯示農村地區的intercept顯著小於4

3.19 The owners of a motel discovered that a defective product was used during construction. It took 7 months to correct the defects during which approximately 14 rooms in the 100-unit motel were taken out of service for 1 month at a time. The data are in the file *motel*.

a. Plot *MOTEL_PCT* and *COMP_PCT* versus *TIME* on the same graph. What can you say about the occupancy rates over time? Do they tend to move together? Which seems to have the higher occupancy rates? Estimate the regression model $MOTEL\_PCT = \beta_1 + \beta_2 COMP\_PCT + e$. Construct a 95% interval estimate for the parameter β_2. Have we estimated the association between *MOTEL_PCT* and *COMP_PCT* relatively precisely, or not? Explain your reasoning.

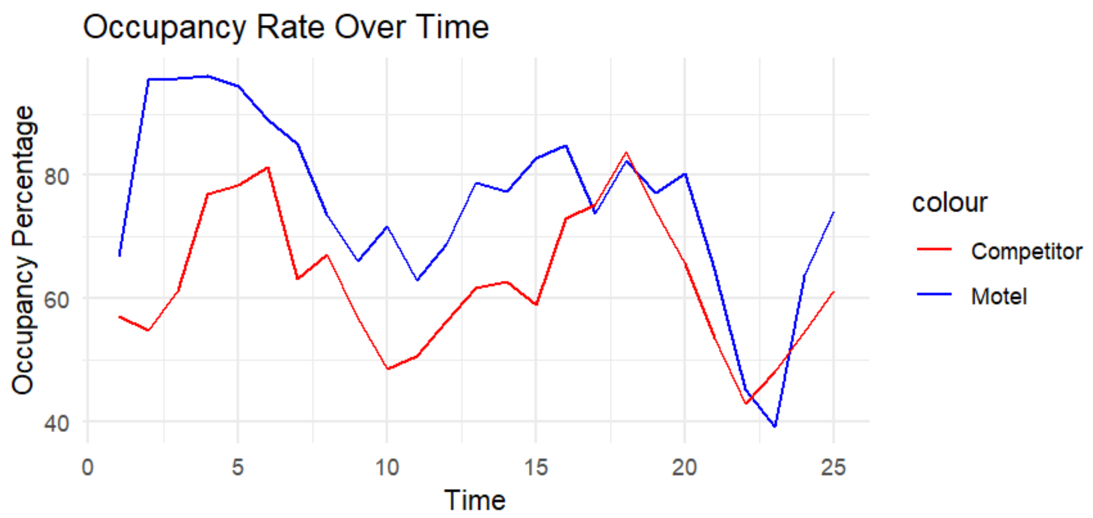
b. Construct a 90% interval estimate of the expected occupancy rate of the motel in question, *MOTEL_PCT*, given that $COMP\_PCT = 70$.

c. In the linear regression model $MOTEL\_PCT = \beta_1 + \beta_2 COMP\_PCT + e$, test the null hypothesis $H_0: \beta_2 \le 0$ against the alternative hypothesis $H_0: \beta_2 > 0$ at the $\alpha = 0.01$ level of significance. Discuss your conclusion. Clearly define the test statistic used and the rejection region.

d. In the linear regression model $MOTEL\_PCT = \beta_1 + \beta_2 COMP\_PCT + e$, test the null hypothesis $H_0: \beta_2 = 1$ against the alternative hypothesis $H_0: \beta_2 \ne 1$ at the $\alpha = 0.01$ level of significance. If the null hypothesis were true, what would that imply about the motel's occupancy rate versus their competitor's occupancy rate? Discuss your conclusion. Clearly define the test statistic used and the rejection region.

e. Calculate the least squares residuals from the regression of *MOTEL_PCT* on *COMP_PCT* and plot them against *TIME*. Are there any unusual features to the plot? What is the predominant sign of the residuals during time periods 17–23 (July, 2004 to January, 2005)?

(a)



兩者之間有相似的變化趨勢，MOTEL_PCT 有較高的 occupancy rates

```
Call:
lm(formula = motel_pct ~ comp_pct, data = motel)

Residuals:
    Min      1Q  Median      3Q     Max 
-23.876  -4.909  -1.193   5.312  26.818 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  21.4000    12.9069   1.658 0.110889    
comp_pct      0.8646     0.2027   4.265 0.000291 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.02 on 23 degrees of freedom
Multiple R-squared:  0.4417,	Adjusted R-squared:  0.4174 
F-statistic: 18.19 on 1 and 23 DF,  p-value: 0.0002906
95% CI: [ 0.4452978 , 1.283981 ]
```

線性回歸估計的結果有落在信賴區間內，顯示我們的估計是可靠的。

(b)

```
       fit      lwr      upr
1 81.92474 77.38223 86.46725
```

(c)

```
> # (c)
> alpha <- 0.01
> t_value <- beta2 / se_beta2
> t_critical <- qt(1 - alpha, df = model$df.residual)
> 
> cat("t 統計量:", t_value)
t 統計量: 4.26536> cat("臨界值:", t_critical)
臨界值: 2.499867>
> if (t_value > t_critical) {
+   cat("拒絕 H0，beta_2 顯著大於 0。")
+ } else {
+   cat("無法拒絕 H0")
+ }
拒絕 H0，beta_2 顯著大於 0。
```

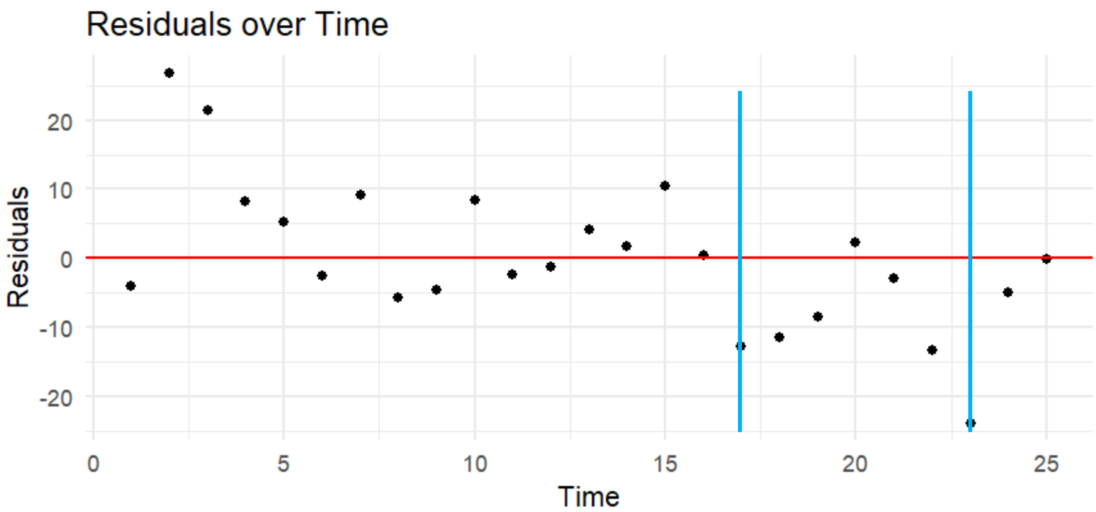
表示 MOTEL_PCT 和 COMP_PCT 之間有正向的關係存在

(d)

```
> # (d)
> t_value_1 <- (beta2 - 1) / se_beta2
> t_critical_1 <- qt(1 - alpha/2, df = model$df.residual)
>
> cat("t 統計量:", t_value_1, "\n")
t 統計量: -0.6677491
> cat("臨界值:", t_critical_1, "\n")
臨界值: 2.807336
>
> if (abs(t_value_1) > t_critical_1) {
+   cat("拒絕 H0，beta_2 與 1 有顯著差異。\n")
+ } else {
+   cat("無法拒絕 H0。\n")
+ }
無法拒絕 H0。
```

表示我們無法拒絕 MOTEL_PCT 和 COMP_PCT 之間的關係為 1：1

(e)



有隨 time 變大，呈現遞減的趨勢

在 time 0~5 之間，殘差多為正，顯示模型低估 MOTEL_PCT 的值

Time 17~23 之間的殘差主要為負，顯示模型高估 MOTEL_PCT 的值