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

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

(b_2 即 $\hat{\beta}_2$；(0.00215) 即 $se(\hat{\beta}_2)$)

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.

單尾

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). $\begin{cases} H_0: \beta_2 = 0 \\ H_1: \beta_2 > 0 \end{cases}$

(b). 使用 t 統計量：$t = \dfrac{\hat{\beta}_2 - 0}{se(\hat{\beta}_2)}$

$\because N = 64$ 若 H_0 成立，則 $t \sim t_{(64-2)} \Rightarrow t \sim t_{(62)}$

(c). 若 H_a 成立，則 $\hat{\beta}_2$ 的期望值會大於 0，t 統計量傾向大於 0，t 分佈的中心將往右偏移。

(d). 顯著水準 1% ($\alpha = 0.01$)，$df = 64 - 2 = 62$

用 R 得 $t_{0.01,62} \approx 2.3880$

```
> qt(0.99, df = 62)
[1] 2.388011
```

拒絕域為 $t > 2.3880$，不拒絕 H_0 的範圍為 $t \le 2.3880$。

(e) $t = \dfrac{\hat{\beta}_2 - 0}{se(\hat{\beta}_2)} = \dfrac{0.01309 - 0}{0.00215} = 6.0884 > 2.3880 \Rightarrow$ 拒絕 H_0

拒絕了 MEDALS 和 GDP 之間沒有關連的假設；不拒絕 H_a，MEDALS 和 GDP 有正向關係

表示在 1% 的顯著水準下，只有 1% 的機率錯誤地拒絕 H_0 (Type I error)，

即誤判 H_a 成立的機率為 1%。

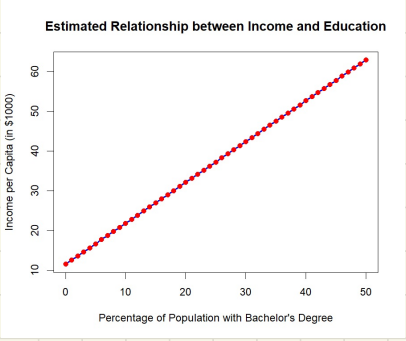
3.7 We have 2008 data on *INCOME* = income per capita (in thousands of dollars) and *BACHELOR* = percentage of the population with a bachelor's degree or more for the 50 U.S. States plus the District of Columbia, a total of $N = 51$ observations. The results from a simple linear regression of *INCOME*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

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

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). $a = \hat{\beta}_1$, $t = \dfrac{a}{se(\hat{\beta}_1)} \Rightarrow 4.31 = \dfrac{a}{2.672} \Rightarrow a = \hat{\beta}_1 \approx 11.5163$ #

(b). INCOME和BACHELOR是以固定速率增加的正向關係。



(c). $C = se(\hat{\beta}_2)$, $t = \dfrac{\beta_2}{C} \Rightarrow 10.75 = \dfrac{1.029}{C} \Rightarrow C = se(\hat{\beta}_2) \approx 0.0957$ #

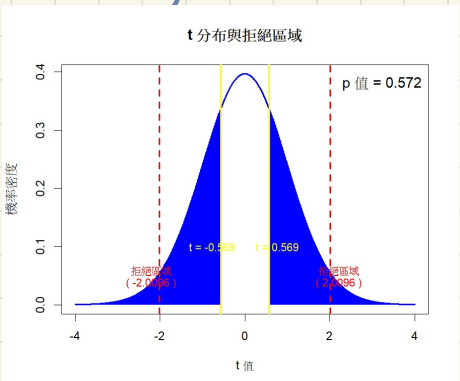
(d). $H_0: \beta_1 = 10$

$$t = \frac{\hat{\beta}_1 - 10}{se(\hat{\beta}_1)} = \frac{11.5163 - 10}{2.672} = 0.5675 \#$$

(e). $H_0: \beta_1 = 10$ 成立的前提，且 $\alpha = 0.05$（雙尾檢定）、$df = 49$ 下

根據題目 $P = 0.572$，遠大於 $\alpha = 0.05$，∴無法拒絕 H_0（出現 H_0 的機率 57.2%）

$\left(P(t > 0.569) = 0.286 \Rightarrow P = 0.286 \times 2 = 0.572\right)$



拒絕域 $t_{\frac{0.05}{2}, 49} = \pm 2.0096$.

藍色部分為 p value.

(f). 99% 信賴區間，$\alpha = 1\%$

$$\hat{\beta}_2 \pm t_{\frac{\alpha}{2}, df} se(\hat{\beta}_2) = 1.029 \pm t_{\frac{0.01}{2}, 49} \times 0.0957$$

用 R 得到 $t_{\frac{0.01}{2}, 49} = 2.6799$

∴ $1.029 \pm 2.6799 \times 0.0957 = (0.7725, 1.2855)$

99% 信心水準下 BACHELOR 對 INCOME 的影響落在這個區間。

(g). $\begin{cases} H_0: \beta_2 = 1 \\ H_1: \beta_2 \neq 1 \text{（雙尾）} \end{cases}$ $\alpha = 0.05$, $df = 49$

$$t = \frac{\hat{\beta}_2 - 1}{se(\hat{\beta}_2)} = \frac{1.029 - 1}{0.0957} = 0.3030$$

用 R 得到 $P(t > 0.3030) \approx 0.382$，又是雙尾 ∴ $0.382 \times 2 = 0.764$

$0.764 > 0.05$ ∴無法拒絕 H_0，BACHELOR 每增加 1% INCOME 增加 $1000.

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.

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

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

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) urban

$H_0: \hat{\beta}_2 = 1.8$

$H_1: \hat{\beta}_2 > 1.8$ (右尾) $\alpha = 0.05$, $df = 986 - 2 = 984$

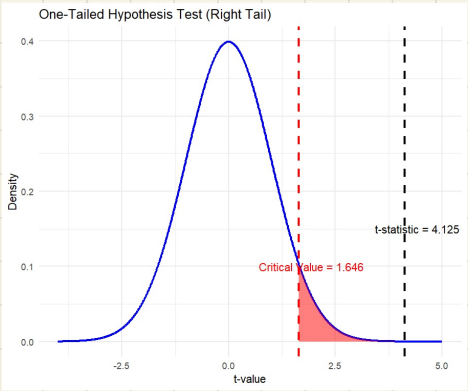
$$t = \frac{\hat{\beta}_2 - 1.8}{se(\hat{\beta}_2)} = \frac{2.46 - 1.8}{0.16} = 4.125$$

拒絕域：$t \sim t_{0.05,984} = 1.6464$

```
> qt(0.95,984)
[1] 1.646404
```

∵ $4.125 > 1.6464$ ∴ 在5%的顯著水準下拒絕 H_0，

即都市地區，EDUC每增加1年對WAGE的影響遠大於1.8 #



3.17

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.

(b) 計算當 EDUC = 16，預測的 WAGE 的 95% 信賴區間

$\alpha = \frac{0.05}{2} = 0.025$

$df = 214 - 2 = 212$

rural: $\widehat{WAGE} = -4.88 + 1.80\,EDUC$

If $EDUC = 16$, $\widehat{WAGE} = -4.88 + 1.80 \times 16 = 23.92$

根據題目 $SE(\widehat{WAGE}) = 0.833$，計算 $\widehat{WAGE}$ 的信賴區間：

$$\widehat{WAGE} \pm t_{0.025,212} \times SE(\widehat{WAGE})$$

其中，$t_{0.025,212}$ 由 R 可得為 1.9712

```
> qt(0.975,212)
[1] 1.971217
```

故 $23.92 \pm 1.9712 \times 0.833 = (22.2780, 25.5620)$

$$SE(\widehat{WAGE}) = \sqrt{1 \times (3.29)^2 + 16^2 \times (0.24)^2 + 2 \times 1 \times 16 \times (-0.761)}$$

$$= 1.1035$$

2.17

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.

(C)、計算當 EDUC = 16，預測的 WAGE 的 95% 信賴區間　$df = 986 - 2 = 984$

$\alpha = \frac{0.05}{2} = 0.025$

Urban：$\widehat{WAGE} = -10.76 + 2.46 \times 16 = 28.6$

$\because se(c_1b_1 + c_2b_2) = \sqrt{c_1^2 \widehat{var}(b_1) + c_2^2 \widehat{var}(b_2) + 2c_1c_2\widehat{cov}(b_1, b_2)}$

$c_1 = 1$，$c_2 = 16$，$cov = -0.345$

$\therefore se(\widehat{WAGE}) = \sqrt{1 \times 2.27^2 + 16^2 \times (0.16)^2 + 2 \times 1 \times 16 \times (-0.345)} = 0.8164$

拖總項 $= t_{0.025, 984} = 1.962378$

```
> qt(0.975,984)
[1] 1.962378
```

Urban 信賴區間：$WAGE \pm t_{\frac{0.05}{2}, 984} \times se(\widehat{WAGE})$

$= 28.6 \pm 1.9624 \times 0.8164 = (26.9979, 30.20)$

Rural 信賴區間：(22.78, 25.56)

Urban 相較於 Rural 區間更寬、標誤更低，可能是 Urban 的樣本數遠大於 Rural，都市教育和工資的關聯變異較小。

3.17

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.

Rural: $\begin{cases} H_0 : \beta_1 \geq 4 \\ H_1 : \beta_1 < 4 \text{ (左尾)} \end{cases}$ $\alpha = 0.01$, $df = 14 - 2 = 12$

$$t = \frac{\hat{\beta}_1 - 4}{se(\hat{\beta}_1)} = \frac{-4.88 - 4}{3.29} = -2.6990$$

拒絕域：$t_{0.01,212} = -2.344$ (左尾)

```
> qt(0.01,212)
[1] -2.344066
```

∵ $-2.699 < -2.344$ ∴拒絕 H_0，

表示當 EDUC=0 時，Urban 的 $\widehat{WAGE}$ 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 $H_0: \beta_2 \leq 0$ 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 $H_0: \beta_2 \neq 1$ 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.

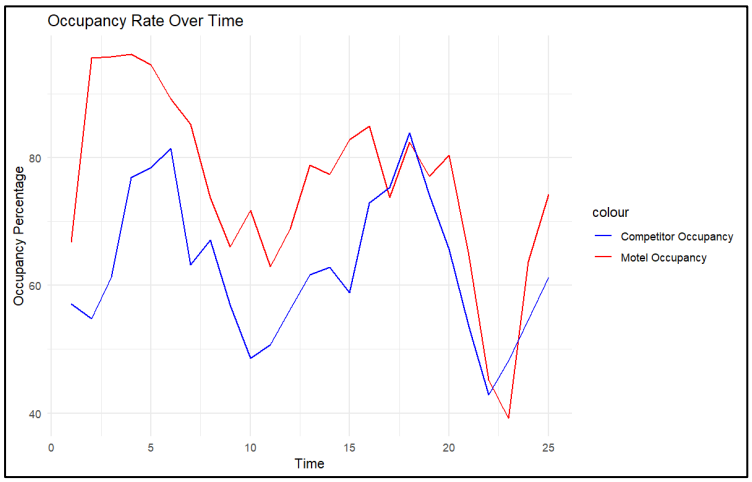
e. Calculate the least squares residuals from the regression of *MOTEL_PCT* on *COMP_PCT* and plot them against *TIME*. Are there any unusual features to the plot? What is the predominant sign of the residuals during time periods 17–23 (July, 2004 to January, 2005)?

(a)

$$\widehat{MOTEL\_PCT} = 21.40 + 0.8646 \cdot COMP\_PCT$$

$$(SE) \qquad (12.9069) \quad (0.2027)$$

在 95%顯著水準下，β_2 的信賴區間為 [0.445, 1.284]。




```
> # 3. 計算 β₂ 的 95% 信賴區間
> confint(model, level = 0.95)
                  2.5 %    97.5 %
(Intercept) -5.2998960 48.099873
comp_pct     0.4452978  1.283981
```

```
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
```

(b)

COMP_PCT = 70下，估算 MOTEL_PCT 的 90%信賴區間為 [0.445, 1.284]

```
> # 計算 90% 信賴區間
> prediction <- predict(model, newdata = new_data, interval = "confidence", level = 0.90)
> print(prediction)
       fit      lwr      upr
1 81.92474 77.38223 86.46725
```

(c)

H_0（虛無假設）：$\beta_2 \leq 0$（競爭者入住率無影響或負影響汽車旅館入住率）

H_1（對立假設）：$\beta_2 > 0$（競爭者入住率正向影響汽車旅館入住率）➔右尾檢定

拒絕 H_0，代表競爭者入住率對汽車旅館入住率有顯著正向影響。

```
> # 計算 t 統計量
> t_value <- beta2 / se_beta2
> print(paste("t 統計量:", t_value))
[1] "t 統計量: 4.26536007134584"
>
> # 計算臨界值（自由度 = n - 2）
> df <- nrow(motel) - 2
> t_critical <- qt(0.99, df)
> print(paste("臨界值 t_critical:", t_critical))
[1] "臨界值 t_critical: 2.49986673949467"
>
> # 計算 p 值（單尾檢定）
> p_value <- 1 - pt(t_value, df)
> print(paste("p 值:", p_value))
[1] "p 值: 0.000145310737114546"
>
> # 檢查是否拒絕 H0
> if (t_value > t_critical) {
+     print("拒絕 H0，競爭者入住率對汽車旅館入住率有顯著正向影響。")
+ } else {
+     print("無法拒絕 H0，競爭者入住率可能無顯著影響。")
+ }
[1] "拒絕 H0，競爭者入住率對汽車旅館入住率有顯著正向影響。"
```

(d)

H_0：$\beta_2 = 1$（汽車旅館入住率與競爭者入住率呈完全正向關係）

H_1：$\beta_2 \neq 1$（汽車旅館入住率與競爭者入住率非完全相同的變動模式）➔雙尾

無法拒絕 H_0，表示汽車旅館入住率的變動模式可能與競爭者入住率一致。

```
> # 計算 t 統計量（測試 β2 是否顯著不同於 1）
> t_value <- (beta2 - 1) / se_beta2
> print(paste("t 統計量:", t_value))
[1] "t 統計量: -0.667749057797476"
>
> # 計算 t 臨界值（雙尾檢定，自由度 = n - 2）
> df <- nrow(motel) - 2
> t_critical <- qt(0.995, df)
> print(paste("臨界值 t_critical:", t_critical))
[1] "臨界值 t_critical: 2.80733568377"
>
> # 計算 p 值（雙尾檢定）
> p_value <- 2 * (1 - pt(abs(t_value), df))
> print(paste("p 值:", p_value))
[1] "p 值: 0.510939188073044"
>
> # 檢查是否拒絕 H0
> if (abs(t_value) > t_critical) {
+     print("拒絕 H0，競爭者入住率與汽車旅館入住率的變動模式不同。")
+ } else {
+     print("無法拒絕 H0，汽車旅館入住率可能與競爭者入住率變動一致（β2 約為 1）。")
+ }
[1] "無法拒絕 H0，汽車旅館入住率可能與競爭者入住率變動一致（β2 約為 1）。"
```

(e)

2004 年 7 月（時間 17）和 2005 年 1 月（時間 23）除了其中一個月，其餘殘差皆為負，表示模型高估了入住率。