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

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

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. $H_0: \beta_2 = 0$

$H_1: \beta_2 > 0$

b.

test statistic : $\frac{b_2 - \beta_2}{SE(b_2)} = \frac{0.01309 - 0}{0.00215} = 6.09$

if the null hypotheses is true $\frac{b_2 - 0}{SE(b_2)} \sim t(n-2) \sim t(62)$

c.

若 H_1 為真，t 統計量期望值變大，分配會往右移

因為正相關意味著 b_2 的估計值大於 0，使 t 值更大

d.

$\alpha = 0.01$ $\quad df = 64 - 2 = 62$

$t_{0.01,62} = 2.388$

If $t^* = \frac{b_2 - 0}{SE(b_2)} > 2.388 = t_{0.01,62}$, reject null hypothesis

If $t^* < 2.338$, donte reject null hypothesis

e.

$t^* = 6.09 > 2.338 = t_{0.01,62}$

∴ reject H_0 有足夠証據說明 GDPB 與獎牌數存在正向關係

經濟解釋：GDP 越高的國家通常能贏得更多的奧運獎牌

這可能是因為經濟好的國家可以投入更多的

資源於運動訓練、基礎設施、科技支援等因素。

1% level of confidence meaning：H_0 為真時，錯誤拒絕的概率為 1%，即結論有 99% 的信心是正確的

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

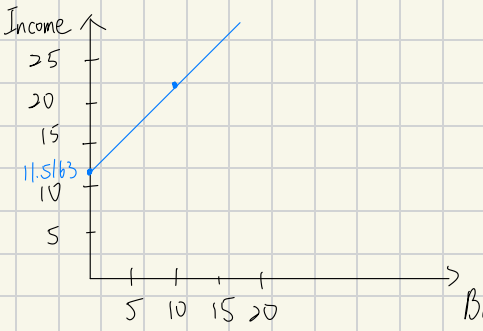
$$\widehat{INCOME} = (a) + 1.029BACHELOR$$
$$\text{se} \quad (2.672) \quad (c)$$
$$t \quad (4.31) \quad (10.75)$$

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.

$$\frac{b_1 - 0}{SE(b_1)} = \frac{a-0}{2.672} = 4.31 \qquad a = 4.31 \times 2.672 = 11.5163 \#$$

b.



斜率為正表示 Income 和 Bachelor 呈正相關

Bachelor 越高，Income 越高，每增加 1% 的 Bachelor，Income 會增加 1029 美元

c.

$$\frac{b_2 - 0}{SE(b_2)} = \frac{1.029 - 0}{c} = 10.75 \qquad c = 0.0957$$

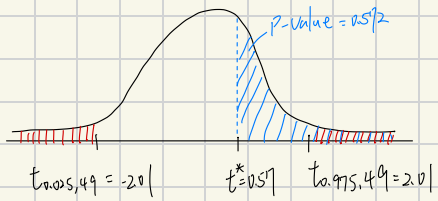
d.

$H_0: \beta_1 = 10$

$$t^* = \frac{b_1 - \beta_1}{SE(b_1)} = \frac{11.5163 - 10}{2.672} = 0.57$$

e.

p-value = 0.572 $\qquad df = 51 - 2 = 49$



$t_{0.025,49} = -2.01$

$t_{0.975,49} = 2.01$

紅色區域為 rejection region

f.

$$b_2 \pm SE(b_2)\, t_{0.005,49} = 1.029 \pm 0.0957 \times 2.68$$

99% interval of slope : $[0.7725, 1.2855]$

g.

$H_0: \beta_2 = 1$

$H_1: \beta_2 \neq 1$

Test statistic : $\dfrac{b_2 - \beta_2}{SE(b_2)} \sim t(49)$

$$t^* = \frac{1.029 - 1}{0.0957} = 0.303 \qquad t_{0.025,49} = 2.01$$

$t^* < t_{0.025,49}$ ∴ don't reject H_0

無足夠證據說明，β_2 顯著不等於 1

經濟解釋：原迴歸表示每增加 1% 的學士學位人口，收入增長約 1.029 千美元，因此可認定 每1%學士學位人口提升，收入增加 1 千美元

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

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

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.

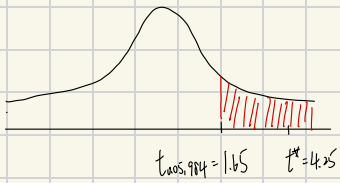
$H_0: \beta_2 = 1.8$

$H_1: \beta_2 > 1.8$

$\alpha = 0.05$

Test statistic : $\frac{b_2 - \beta_2}{SE(b_2)} \sim t(984)$

$t^* = \frac{2.46 - 1.8}{0.16} = 4.125 \qquad t_{0.05, 984} = 1.65$



$t^* > t$, t^* 在拒絕域中, ∴ reject H_0

有足夠証據說明斜率 Urban 迴歸式斜率大於 1.8，表示當 EDUC 增加一個單位(年) WAGE 增加 1 美元

b.

$\widehat{WAGE} = -4.88 + 1.8 \times 16 = 23.92$

EDUC = 16, WAGE 的 95% interval

$23.92 \pm 0.833 \times t_{0.025, 212} = 23.92 \pm 0.833 \times 1.97$

$[22.278, 25.562]$

$SE(b_1 + b_2 \times 16) = \sqrt{Var(b_1 + b_2 \times 16)}$

$= \sqrt{Var(b_1) + 16^2 Var(b_2) + 2 \times 16 \times cov(b_1, b_2)}$

$= \sqrt{3.29^2 + 16^2 \times .024^2 + 2 \times 16 \times (-0.761)}$

$= 1.10349$

c.

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

$SE(\widehat{WAGE}) = \sqrt{Var(b_1 + b_2 16)}$

$= \sqrt{Var(b_1) + 16^2 Var(b_2) + 2 \times 16 \times cov(b_1, b_2)}$

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

$= 0.8164$

$28.6 \pm 0.8164\, t_{0.025, 984} = 28.6 \pm 0.8164 \times 1.96$

$\widehat{WAGE}$ 的 95% interval $[26.998, 30.202]$

Urban 的信賴區間相較 Rural 更窄，這是合理的，因為 Urban 的樣本較大 誤差會更小

d.

$H_0: \beta_1 \geq 4$

$H_1: \beta_1 < 4$

$\alpha = 0.01$

Test statistic : $\frac{b_1 - 4}{SE(b_1)} \sim t(214-2)$

$t^* = \frac{-4.88 - 4}{3.29} = -2.699 \qquad t_{0.01, 212} = -2.34$

$t^* < t_{0.01, 212}$, ∴ reject H_0

表示有足夠証據說明 $\beta_1 < 4$，也就是當 EDUC 為 0 WAGE 會小於 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.

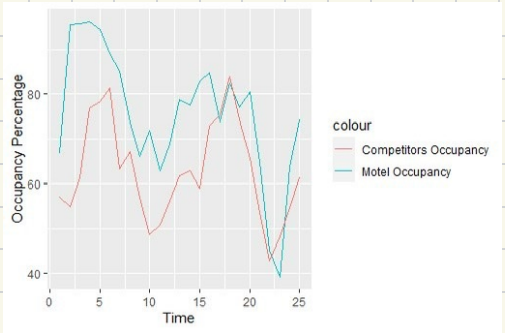
b. Construct a 90% interval estimate of the expected occupancy rate of the motel in question, *MOTEL_PCT*, given that $COMP\_PCT = 70$.

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.



Motel 和 Competitors 的入住率都隨時間變化有一定的波動

兩條線的變動趨勢相似

大概在 Time=17 以前 Motel Occupancy rate 較高

在 Time=17 之後，Motel 和 Competitors 各有高低

```
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
```

$\widehat{MOTEL\_PCT} = 21.4 + 0.8646\, COMP\_PCT$

```
> confint(m1,level=0.95)
                2.5 %    97.5 %
(Intercept) -5.2998960 48.099873
comp_pct     0.4452978  1.283981
```

β_2 95% interval : $b_2 \pm SE(b_2)\, t_{(25-2)}$

$[0.4453, 1.284]$

從 summary 的報表中可以看到係數 β_2 的 p-value 為 0.00029 可知在 $\alpha = 0.001$ 下顯著

由此可知 β_2 是有解釋能力的

b.

```
> predict(m1,new_data,interval = 'confidence',level=0.9)
       fit      lwr      upr
1 81.92474 77.38223 86.46725
```

當 $COMP\_PCT = 70$ 時，$\widehat{MOTEL\_PCT} = 81.9247$

MOTEL_PCT 90% interval 為 $[77.3822, 86.4673]$

c.

$H_0: \beta_2 \le 0$

$H_1: \beta_2 > 0$

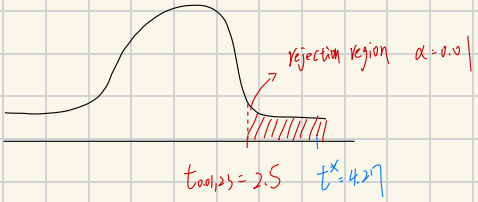
$\alpha = 0.05$

$t = \frac{b_2 - 0}{SE(b_2)} \sim t_{(25-2)}$

$t^* = \frac{0.865 - 0}{0.2027} = 4.27 \qquad t_{0.01,23} = 2.5$

p-value = 0.00015 < 0.01 , ∴ reject H_0

有足夠證據說明 $\beta_2 > 0$ 表示 COMP_PCT 與 MOTEL_PCT 呈正相關



t^* 在 reject region 中

d.

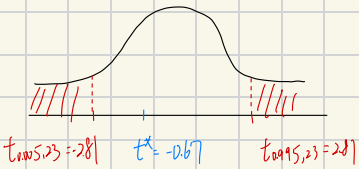
$H_0: \beta_2 = 1$

$H_1: \beta_2 \ne 1$

$\alpha = 0.01$

$t = \frac{b_2 - 1}{SE(b_2)} \sim t_{(25-2)}$

$t^* = \frac{b_2 - 1}{SE(b_2)} = -0.67 \qquad t_{0.005,23} = -2.81$

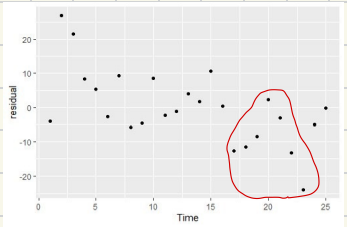


t^* 未在 rejection region 中，∴ don't reject H_0

無足夠證據說明 $\beta_2 \ne 1$，也就是說當 COMP_PCT 增加 1 個單位(%)，會使 MOTEL_PCT 增加 1 個單位(%)

這也間接證明 COMP_PCT 和 MOTEL_PCT 變動幅度差不多一樣

e.



SSE = 2792.52

The residuals are very large during time periods 2.3 and 23

The occupancy rate of the motel is higher than competitor's occupancy rate.