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

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

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.

$$H_0: \beta_2 = 0$$
$$H_1: \beta_2 > 0$$

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

$$t = \frac{b_2 - 0}{se(b_2)} = \frac{0.01309}{0.00215} = 6.088372$$

distribution 為自由度=62 的 t 分配

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?]

若H_1為真，則$\beta_2 > 0$,分配將會向右移(因為$E(\beta_2) > 0$)

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?

自由度 62, $\alpha = 0.012$ 的 t 值約為 2.39

若檢驗結果$t > 2.39$則拒絕H_0, 若$t \leq 2.39$則不拒絕H_0

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?

$t = 6.088372 > 2.39$➔拒絕H_0，表示 GDP 與獎牌數之間呈正向關係，1%顯著水準意味著拒絕H_0的機率僅有 1%，而在此拒絕表示對於 GDP 與獎牌數之間呈正向關係有 99%的信心。

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

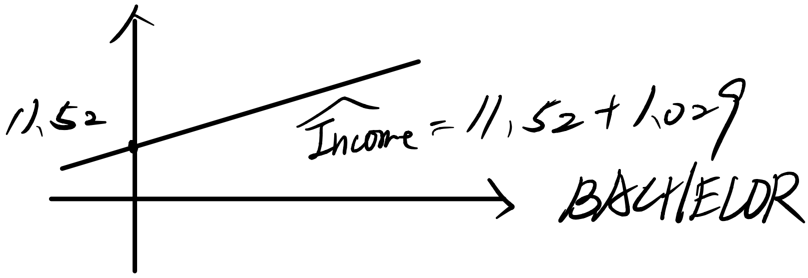
$$\begin{array}{rcccc} \widehat{INCOME} = & (a) & + & 1.029BACHELOR \\ \text{se} & (2.672) & & (c) \\ t & (4.31) & & (10.75) \end{array}$$

a. Using the information provided calculate the estimated intercept. Show your work.

$$t = \frac{estimated\ intercept}{se}, 4.31 = \frac{a}{2.672}, a = 11.51632$$

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?

Increasing
positive relationship
increasing at a constant rate



c. Using the information provided calculate the standard error of the slope coefficient. Show your work.

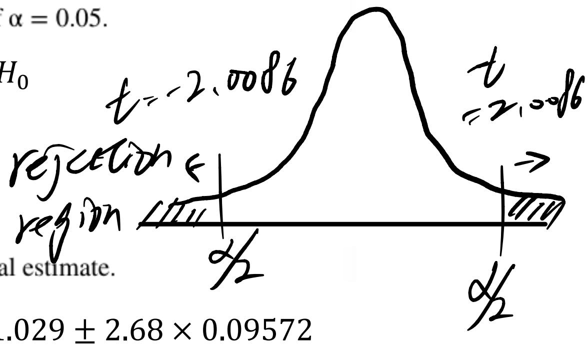
$$10.75 = \frac{1.029}{se}, se = 0.09572$$

d. What is the value of the t-statistic for the null hypothesis that the intercept parameter equals 10?

$$t = \frac{a - 10}{se(a)} = \frac{11.52 - 10}{2.672} = 0.5689$$

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 $\alpha = 0.05$.

$\alpha = 0.05, t_{0.025,50} = 2.0086, 0.5689 < 2.0086$ ➔不拒絕H_0



f. Construct a 99% interval estimate of the slope. Interpret the interval estimate.

$$1.029 \pm t_{\frac{\alpha}{2},df} \times se = 1.029 \pm t_{0.005,49} \times 0.09572 = 1.029 \pm 2.68 \times 0.09572$$

Interval:(0.7725,1.2855)，有 99%信心真實的斜率會落在此區間中

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.

$$H_0: \beta_1 = 1, H_1: \beta_1 \neq 1$$

$$t_{0.025,50} = -2.0086 < t = \frac{\beta_1 - 1}{se(\beta_1)} = \frac{1.029 - 1}{0.09572} = 0.30296 < t_{0.025,50} = 2.0086$$

➔不拒絕H_0

➔無法拒絕斜率=1 的假設，在斜率=1 之下，每增加 1%BACHELOR 對 INCOME 的影響是增加 1000 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.

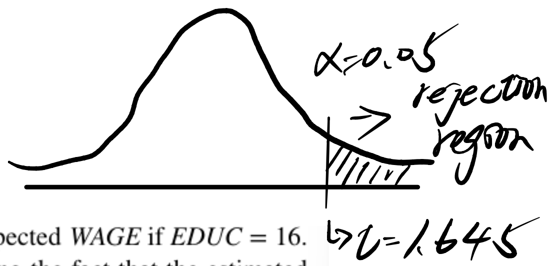
$$\text{Urban} \quad \widehat{WAGE} = \underset{(2.27)}{-10.76} + \underset{(0.16)}{2.46}EDUC, \quad N = 986 \quad \text{(se)}$$

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

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.

$H_0: \beta_2 = 1.8, H_1: \beta_2 > 1.8$

$t = \frac{\beta_2 - 1.8}{se(\beta_2)} = \frac{2.46 - 1.8}{0.16} = 4.125$



For one tail $t_{0.05,984} = 1.645, 4.125 > 1.645$➔拒絕$H_0$

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.

$$E(WAGE) = -4.88 + 1.8 \times 16 = 23.92$$

$$se(WAGE) = \sqrt{3.29^2 + 16^2 \times 0.24^2 + 2 \times 16 \times (-0.761)} = 1.1035$$

$$23.92 \pm t_{0.025,212} \times 1.1035 = 23.95 \pm 1.971 \times 1.1035$$

CI:(21.745,26.125)

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.

$$E(WAGE) = -10.76 + 2.46 \times 16 = 28.6$$

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

$$28.6 \pm t_{0.025,984} \times 0.8164 = 28.6 \pm 1.961 \times 0.8164$$

CI:(26.999,30.2)

Rural 區間範圍約為 4.35, urban 約為 3.2，較窄，主要是因在 urban 的標準差較小，使得同樣信心水準下的區間範圍較小

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.

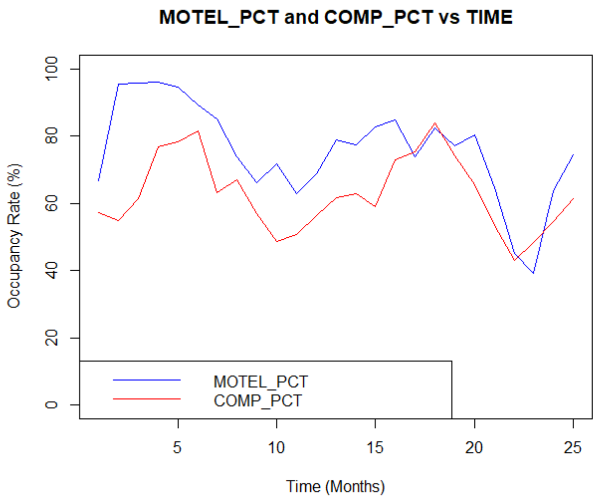
$$H_0: \beta_1 = 4, H_1: \beta_1 < 4$$

$$t = \frac{\beta_1 - 4}{se(\beta_1)} = \frac{-4.88 - 4}{3.29} = -2.699$$

For left tail $t_{0.01,212} = -2.33, -2.699 < -2.33 \rightarrow$ 拒絕H_0

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.



Motel 有較高的 occupancy rates

Tend to move together

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

$$MOTEL_{PCT} = 21.4 + 0.8646 \times COMP\_PCT$$

```
> # 95% confidence interval for beta2
> confint(model, "motel$comp_pct" , level = 0.95)
                   2.5 %   97.5 %
motel$comp_pct 0.4452978 1.283981
```

CI: (0.4453,1.284)

Relatively precisely

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

```
> #b.
> # Predict MOTEL_PCT with a 90% confidence interval for the mean
> beta1 <- coef(model)[1]  # 截距
> beta2 <- coef(model)[2]  # 斜率
>
> # 計算預測值 (COMP_PCT = 70)
> comp_pct_new <- 70
> motel_pct_pred <- beta1 + beta2 * comp_pct_new
>
> n <- nrow(motel)  # 樣本量
>
> x_mean <- mean(motel$comp_pct)  # COMP_PCT 的均值和離差平方和
> x_ss <- sum((motel$comp_pct - x_mean)^2)
> sigma <- summary(model)$sigma  # 模型的殘差標準誤
> se_yhat <- sqrt(sigma^2 * (1/n + (comp_pct_new - x_mean)^2 / x_ss)) # 預測值的標準誤
>
> ci_lower <- motel_pct_pred - qt(0.95, n - 2) * se_yhat
> ci_upper <- motel_pct_pred + qt(0.95, n - 2) * se_yhat
> cat(sprintf("90%% Confidence Interval: (%.2f, %.2f)\n", ci_lower, ci_upper))
90% Confidence Interval: (77.38, 86.47)
```

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.

```
> cat("Test Statistic (t):", t_stat_c, "\n")
Test Statistic (t): 4.26536
```

```
> cat("Critical Value (t at alpha = 0.01, df =", df, "):", t_c, "\n")
Critical Value (t at alpha = 0.01, df = 23 ): 2.499867
```

```
> cat("Rejection Region: t >", t_c, "\n")  # 拒絕域
Rejection Region: t > 2.499867
```

```
+ }  # 檢定結論
Reject H0 at alpha = 0.01.
This suggests a positive relationship between COMP_PCT and MOTEL_PCT.
>
```

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.

```
> cat("Test Statistic (t):", t_stat_d, "\n")
Test Statistic (t): -0.6677491
```

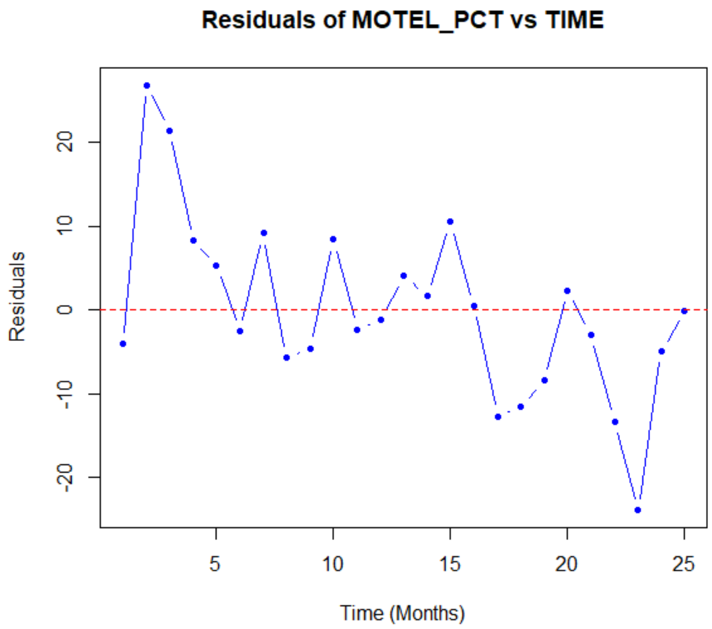
```
> cat("Critical Values (t_critical at alpha = 0.01, df =", df, "): ±", t_d, "\n")
Critical Values (t_critical at alpha = 0.01, df = 23 ): ± 2.807336
```

```
> cat("Rejection Region: t <", -t_d, "or t >", t_d, "\n")  # 拒絕域
Rejection Region: t < -2.807336 or t > 2.807336
```

```
+ }  # 檢定結論
Fail to reject H0 at alpha = 0.01.
```

表示 COMP_PCT 上升一單位，MOTEL_PCT 也上升一單位

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)?




```
> print(data.frame(Time = motel$time[17:23], Residuals = residuals_17_23,
   Time  Residuals Sign
17   17 -12.707328   -1
18   18 -11.543226   -1
19   19  -8.456225   -1
20   20   2.279673    1
21   21  -2.958191   -1
22   22 -13.293015   -1
23   23 -23.875603   -1
> cat("Predominant sign of residuals during time periods 17-23:", predomi
"\n")
Predominant sign of residuals during time periods 17-23: Negative
```

Residual 看起來不是隨機的(在 17-23 大多是負的)，因此模型可能未考量到某些因子