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

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

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.

$H_0: b_2 = 0$, $H_1: b_2 > 0$ *

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

$$t = \frac{0.01309}{0.00215} \approx 6.088$$

if H_0 is true $\Rightarrow$ 自由度 $= 64 - 2 = 62$

$\Rightarrow t \sim t(62)$ *

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?]

If H_1 is true, the t-statistic follows a t-distribution with n-2 degrees of freedom and has a mean of 0. However, if H_1 is true, the mean of the t-statistic increases, causing the distribution to shift to the right. As a result, when H_1 is true, we are more likely to observe larger t-values, increasing the probability of rejecting H_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?

```
> # 顯示結果
> t_statistic
[1] 6.088372
> t_critical
[1] 2.388011
```

$\because 2.388011 < 6.088372$

$\Rightarrow$ reject 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?

```
> # 顯示結果
> p_value
[1] 3.943571e-08
```

Hypothesis Testing Conclusion:

1. $\because t = 6.088 >$ the critical value 2.388, reject H_0 at the 1% significance lael.

2. Checking the p-value:

   The p-value should be very small, indicating that the result is highly significant.

Economic implications:

1. GDP has a significant positive impact on the number of Olympic medals won.

2. The meaning of the 1% singnificance lael: We are willing to accept at must a 1% probability of mistakenly rejecting H_0.

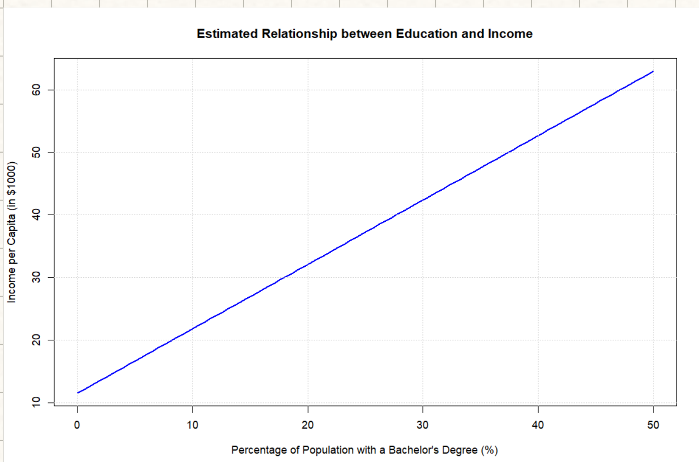
3.7 We have 2008 data on *INCOME* = income per capita (in thousands of dollars) and *BACHELOR* = percentage of the population with a bachelor's degree or more for the 50 U.S. States plus the District of Columbia, a total of $N = 51$ observations. The results from a simple linear regression of *INCOME*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

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

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

$\hat{a} = t \cdot se(\hat{a}) = 4.31 \times 2.672 = 11.51632$

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?



Slope $= 1.029 > 0 \Rightarrow$ increasing

$\because 1.029 > 0 \Rightarrow$ positive

it increasing at a constant rate

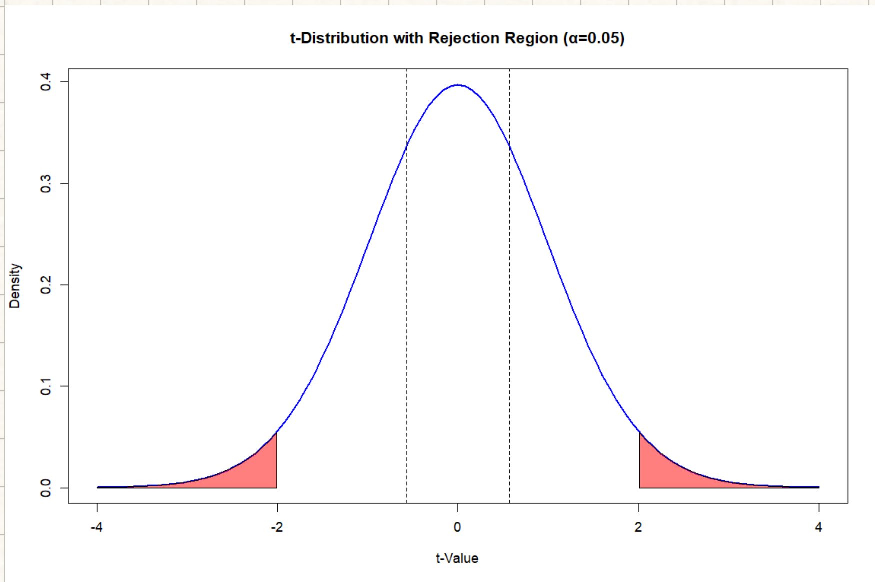
c. Using the information provided calculate the standard error of the slope coefficient. Show your work.

$t = \frac{\text{slope}}{(c)} \Rightarrow (c) = \frac{\text{slope}}{t} = \frac{1.029}{10.75} = 0.09572$

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

$t = \frac{\hat{a} - a}{se(\hat{a})} = \frac{11.51632 - 10}{2.672} = 0.5673653$

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.

```
> # 已知數值
> slope_estimate <- 1.029  # 斜率估計值
> se_slope <- 0.0957  # 斜率標準誤
> df <- 50  # 自由度
> confidence_level <- 0.99  # 99% 信賴區間
>
> # 計算 t 臨界值 (99% 信賴區間，雙尾)
> t_critical_99 <- qt(1 - (1 - confidence_level) / 2, df)
>
> # 計算信賴區間
> lower_bound <- slope_estimate - t_critical_99 * se_slope
> upper_bound <- slope_estimate + t_critical_99 * se_slope
>
> # 顯示結果
> c(lower_bound, upper_bound)
[1] 0.7727352 1.2852648
```

$\Rightarrow (0.7727352, 1.2852648)$

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: slope $= 1$, H_1: slope $\neq 1$

```
> # 已知數值
> slope_estimate <- 1.029  # 斜率估計值
> null_hypothesis_slope <- 1  # H0: 斜率 = 1
> se_slope <- 0.0957  # 斜率標準誤
> df <- 50  # 自由度
> alpha <- 0.05  # 顯著水準
>
> # 計算 t-統計量
> t_statistic_slope <- (slope_estimate - null_hypothesis_slope) / se_slope
>
> # 計算 p-value (雙尾檢定)
> p_value_slope <- 2 * (1 - pt(abs(t_statistic_slope), df))
>
> # 計算 5% 顯著水準的 t-臨界值
> t_critical_95 <- qt(1 - alpha/2, df)
>
> # 檢查是否拒絕 H0
> reject_null <- abs(t_statistic_slope) > t_critical_95
>
> # 顯示結果
> c(t_statistic_slope, p_value_slope, t_critical_95, reject_null)
[1] 0.3030303 0.7631240 2.0085591 0.0000000
```

t-statistic: 0.303

p-value: 0.763

critical value of t: ± 2.009

$\because$ $0.303 < 2.009$ and $0.763 > 0.05$

$\Rightarrow$ Conclusion: not reject H_0

Economic result:

the relationship between a bachelor's degree and income may follow a 1:1 linear relationship.

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.

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

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

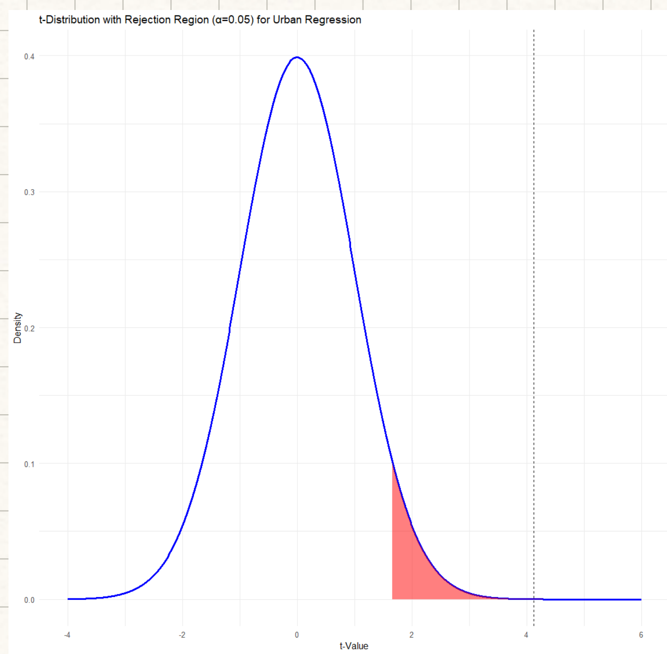
a. Using the urban regression, test the null hypothesis that the regression slope equals 1.80 against the alternative that it is greater than 1.80. Use the $\alpha = 0.05$ level of significance. Show all steps, including a graph of the critical region and state your conclusion.

$H_0: \beta_2 = 1.80, \quad H_1: \beta_2 > 1.80$

$t = \dfrac{2.46-1.80}{0.16} = 4.125$, 自由度 $= 986-2 = 984$

Critical value of t under $\alpha = 0.05$ level of significance $= 1.646$

```
> # 載入必要的套件
> library(ggplot2)
>
> # 已知數值
> slope_estimate_urban <- 2.46  # 斜率估計值
> null_hypothesis_slope <- 1.80  # H0: 斜率 = 1.80
> se_slope_urban <- 0.16  # 斜率標準誤
> df_urban <- 986 - 2  # 自由度 (N-2)
> alpha <- 0.05  # 顯著水準
>
> # 計算 t-統計量
> t_statistic_urban <- (slope_estimate_urban - null_hypothesis_slope) / se_slope_urban
>
> # 計算 p-value (右尾檢定)
> p_value_urban <- 1 - pt(t_statistic_urban, df_urban)
>
> # 計算 5% 顯著水準的 t-臨界值 (右尾檢定)
> t_critical_95_urban <- qt(1 - alpha, df_urban)
>
> # 建立 t-分佈數據
> x_vals <- seq(-4, 6, length.out = 1000)
> y_vals <- dt(x_vals, df_urban)
>
> # 建立 DataFrame
> data <- data.frame(x = x_vals, y = y_vals)
>
> # 繪製 t-分佈曲線
> p <- ggplot(data, aes(x, y)) +
+     geom_line(color = "blue", linewidth = 1) +  # 修正這裡
+     geom_area(data = subset(data, x >= t_critical_95_urban),
+               aes(x, y), fill = "red", alpha = 0.5) +  # 拒絕區域
+     geom_vline(xintercept = t_statistic_urban, color = "black", linetype = "dashed") +  # 標示 t-統計量
+     labs(title = "t-Distribution with Rejection Region (α=0.05) for Urban Regression",
+          x = "t-Value", y = "Density") +
+     theme_minimal()
>
> # 顯示圖表
> print(p)
>
> # 顯示檢定結果
> c(t_statistic_urban, p_value_urban, t_critical_95_urban, t_statistic_urban > t_critical_95_urban)
[1] 4.125000e+00 2.010319e-05 1.646404e+00 1.000000e+00
```



$\Rightarrow$ since t-statistic $= 4.125 >$ critical value $= 1.646 \Rightarrow$ Reject H_0 #

$\Rightarrow$ the impact of education on wages in urban areas is significantly greater than 1.8. #

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.

$$\widehat{WAGE} = (-4.88) + (1.8 \times 1.6) = 23.92$$

$$t_{0.975/212} = 1.971$$

$$CI = \widehat{WAGE} \pm t_{\frac{\alpha}{2}, df} \times SE(\widehat{WAGE}) = 23.92 \pm (1.971 \times 0.833)$$

$$= (22.278, 25.562)$$

*

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.

```
> # 已知數值
> intercept_urban <- -10.76  # 截距
> slope_urban <- 2.46  # 斜率
> educ_value <- 16  # EDUC = 16
> se_intercept_urban <- 2.27  # 截距標準誤
> se_slope_urban <- 0.16  # 斜率標準誤
> cov_intercept_slope_urban <- -0.345  # 截距與斜率共變異數
> df_urban <- 986 - 2  # 自由度
> alpha <- 0.05  # 顯著水準
>
> # 計算預測工資
> wage_estimate_urban <- intercept_urban + slope_urban * educ_value
>
> # 計算標準誤
> se_wage_urban <- sqrt(se_intercept_urban^2 + (educ_value^2) * (se_slope_urban^2) + 2 * educ_value * cov_intercept_slope_urban)
>
> # 計算 95% 信賴區間
> t_critical_95_urban <- qt(1 - alpha / 2, df_urban)
> lower_bound_urban <- wage_estimate_urban - t_critical_95_urban * se_wage_urban
> upper_bound_urban <- wage_estimate_urban + t_critical_95_urban * se_wage_urban
>
> # 顯示結果
> c(wage_estimate_urban, se_wage_urban, lower_bound_urban, upper_bound_urban)
[1] 28.6000000  0.8163945 26.9979256 30.2020744
```

$\widehat{WAGE} = 28.6$

$\Rightarrow SE(\widehat{WAGE}) = 0.816$

95% CI : (27.00, 30.20)

The CI for urban regression is narrower than that for rural regression.

This is because the urban sample size is larger ($N = 986$), leading to less variability in the estimates, this is reasonable. *

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.

```
> # 已知數值
> intercept_rural <- -4.88  # 截距估計值
> null_hypothesis_intercept <- 4  # H0: 截距 = 4
> se_intercept_rural <- 3.29  # 截距標準誤
> df_rural <- 214 - 2  # 自由度
> alpha <- 0.01  # 顯著水準 (1%)
>
> # 計算 t-統計量
> t_statistic_intercept_rural <- (intercept_rural - null_hypothesis_intercept) / se_intercept_rural
>
> # 計算 p-value (左尾檢定)
> p_value_intercept_rural <- pt(t_statistic_intercept_rural, df_rural)
>
> # 計算 1% 顯著水準的 t-臨界值 (左尾檢定)
> t_critical_99_rural <- qt(alpha, df_rural)
>
> # 檢查是否拒絕 H0
> reject_null_intercept_rural <- t_statistic_intercept_rural < t_critical_99_rural
>
> # 顯示結果
> c(t_statistic_intercept_rural, p_value_intercept_rural, t_critical_99_rural, reject_null_intercept_rural)
[1] -2.699088146  0.003756828 -2.344065819  1.000000000
```

$H_0: \beta_1 \geq 4, \quad H_1: \beta_1 < 4$

$\Rightarrow$ t-statistic = −2.7

p-value = 0.00376

critical value of $t(\alpha = 0.01) = -2.344$

$\Rightarrow$ Conclusion: Not reject H_0

$\Rightarrow$ The low-education population in rural areas has a baseline wage significantly lower than \$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.

```
> # 8 顯示回歸結果
> summary(model)

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
```

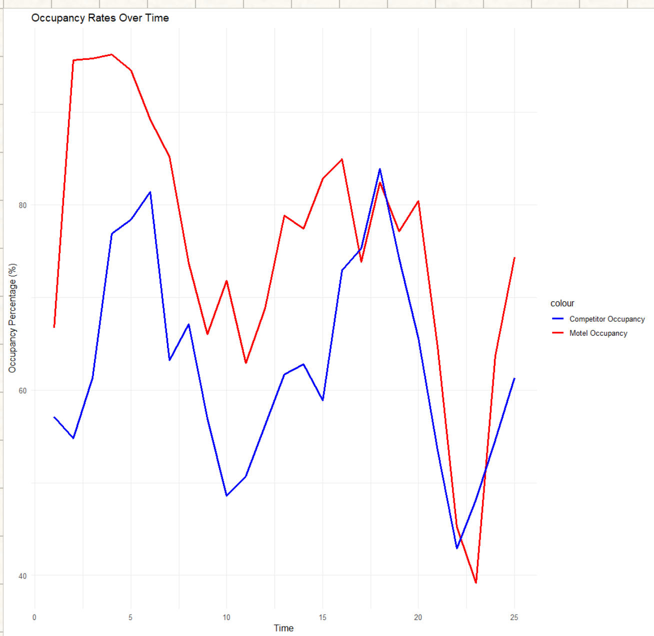
```
> # 10 顯示 95% 信賴區間
> c(lower_bound, upper_bound)
 comp_pct  comp_pct
0.4452978 1.2839809
```

$\Rightarrow$ Motel_PCT $= 21.4 + 0.8646 \times$ COMP_PCT

$R^2 = 0.4417$

p-value $= 0.0002906$

$\Rightarrow$ 95% CI: $(0.4452978, 1.2839809)$



1. The occupancy rates fluctuate over time

2. They tend to have a positive relationship, so tend to move together.

3. The Motel has a higher occupany rates.

4. The estimated association between MOTEL_PCT and COMP_PCT is statistically significant. However, the 95% CI is relatively wide, suggest that some level of uncertainly in the precision of our estimate. #

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

```
> # Given COMP_PCT = 70
> comp_pct_value <- 70
>
> # Predict the expected occupancy rate (MOTEL_PCT) at COMP_PCT = 70
> predicted_motel_pct <- coef(model)[1] + coef(model)[2] * comp_pct_value
>
> #  Calculate the standard error of the estimate (SE_Yhat)
> n <- nrow(motel)  # Number of observations
> df <- n - 2  # Degrees of freedom
> t_critical_90 <- qt(1 - 0.10 / 2, df)  # t-value for 90% confidence leve
l
>
> # Extract residual standard error (sigma hat)
> sigma_hat <- summary(model)$sigma
>
> # Compute SE_Yhat (Standard Error of the Prediction)
> x_mean <- mean(motel$comp_pct)
> se_Yhat <- sigma_hat * sqrt(1/n + (comp_pct_value - x_mean)^2 / sum((mot
el$comp_pct - x_mean)^2))
>
> # Compute the confidence interval
> lower_bound <- predicted_motel_pct - t_critical_90 * se_Yhat
> upper_bound <- predicted_motel_pct + t_critical_90 * se_Yhat
>
> # Display the results
> cat("90% Confidence Interval for MOTEL_PCT when COMP_PCT = 70:\n")
90% Confidence Interval for MOTEL_PCT when COMP_PCT = 70:
> c(lower_bound, upper_bound)
(Intercept) (Intercept)
   77.38223    86.46725
```

$\Rightarrow$ 90% CI

$= (77.38223, 86.46725)$ #

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.

```
> # 提取回歸係數與標準誤
> beta2_hat <- coef(model)[2]  # 估計的斜率
> se_beta2 <- summary(model)$coefficients[2, 2]  # 斜率的標準誤
>
> # 計算 t 統計量
> t_stat <- beta2_hat / se_beta2
>
> # 計算臨界值 (α = 0.01，右尾檢定)
> alpha <- 0.01
> df <- nrow(motel) - 2  # 自由度
> t_critical <- qt(1 - alpha, df)  # 右尾臨界值
>
> # 顯示結果
> cat("檢定統計量 (t):", t_stat, "\n")
檢定統計量 (t): 4.26536
> cat("臨界值 (t_alpha=0.01):", t_critical, "\n")
臨界值 (t_alpha=0.01): 2.499867
>
> # 判斷是否拒絕 H0
> if (t_stat > t_critical) {
+     cat("結論：拒絕 H0，表示 COMP_PCT 顯著影響 MOTEL_PCT。\n")
+ } else {
+     cat("結論：無法拒絕 H0，無法確認 COMP_PCT 對 MOTEL_PCT 有顯著影響。\n")
+ }
結論：拒絕 H0，表示 COMP_PCT 顯著影響 MOTEL_PCT。
```

#

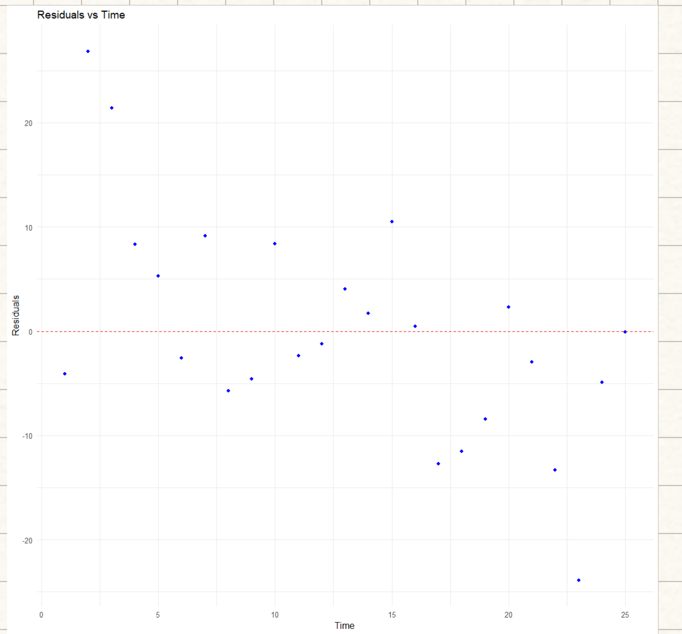
d. In the linear regression model $MOTEL\_PCT = \beta_1 + \beta_2 COMP\_PCT + e$, test the null hypothesis $H_0 : \beta_2 = 1$ against the alternative hypothesis $H_0 : \beta_2 \neq 1$ at the $\alpha = 0.01$ level of significance. If the null hypothesis were true, what would that imply about the motel's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

```
> # 提取回歸係數與標準誤
> beta2_hat <- coef(model)[2]  # 估計的斜率
> se_beta2 <- summary(model)$coefficients[2, 2]  # 斜率的標準誤
>
> # 計算 t 統計量（假設 H0: β2 = 1）
> t_stat <- (beta2_hat - 1) / se_beta2
>
> # 計算臨界值（α = 0.01，雙尾檢定）
> alpha <- 0.01
> df <- nrow(motel) - 2  # 自由度
> t_critical <- qt(1 - alpha/2, df)  # 雙尾檢定的臨界值
>
> # 顯示結果
> cat("檢定統計量 (t):", t_stat, "\n")
檢定統計量 (t): -0.6677491
> cat("臨界值 (t_alpha=0.01, 雙尾):", t_critical, "\n")
臨界值 (t_alpha=0.01, 雙尾): 2.807336
>
> # 判斷是否拒絕 H0
> if (abs(t_stat) > t_critical) {
+     cat("結論：拒絕 H0，表示 β2 不等於 1，競爭者入住率與旅館入住率變動幅度不同。\n")
+ } else {
+     cat("結論：無法拒絕 H0，表示 β2 可能等於 1，競爭者入住率與旅館入住率變動幅度相同。\n")
+ }
結論：無法拒絕 H0，表示 β2 可能等於 1，競爭者入住率與旅館入住率變動幅度相同。
```

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)?

```
> # 1 計算 OLS 殘差
> motel$residuals <- residuals(model)  # 從回歸模型提取殘差
>
> # 2 繪製殘差圖
> library(ggplot2)
>
> ggplot(motel, aes(x = time, y = residuals)) +
+     geom_point(color = "blue") +  # 繪製殘差點
+     geom_hline(yintercept = 0, linetype = "dashed", color = "red") +  #
水平線表示殘差 = 0
+     labs(title = "Residuals vs Time",
+          x = "Time",
+          y = "Residuals") +
+     theme_minimal()
>
> # 3 檢查時間區間 17-23 內的殘差符號
> subset_residuals <- motel$residuals[motel$time >= 17 & motel$time <= 23]
> table(sign(subset_residuals))  # 統計正負殘差的數量

-1  1
 6  1
```



⇒ Residual Plot Observations :

The residual plot shows systematic variation rather than pure randomness,

suggesting : 1. Possible seasonality

2. Residuals tend to be positive in some periods and negative in others, indicating model underestimation or overestimation at different time.

⇒ In time Periods 17-23 :

There are 1 positive residual and 6 negative residual.

This means that during July 2004 - January 2005, the motel's actual occupany was lower than predicted.

*