# HW0310

Yung-Jung Cheng

2025-03-15

# Q1

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

$$ME\hat{D}ALS = b_1 + b_2 GDPB = 7.61733 + 0.01309 GDPB$$

# Q1 (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.

## Ans

- $H_0 : b_2 = 0$
- $H_1 : b_2 > 0$

# Q1 (b)

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

## Ans

- $t = \frac{0.01309-0}{0.00215} \approx 6.088$
- the distribution will be $t_{64-2} = t_{62}$ if the null hypothesis is true.

# Q1 (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 b2 if the null hypothesis is true, and what is it if the alternative is true?

## Ans

If the alternative hypothesis is true, the distribution will shift to **right**, indicating that positive t-values are more likely to occur. This reflects a positive correlation between GDP and the number of medals won.

# Q1 (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?

## Ans

```r
# 定義參數
b2_estimate <- 0.01309  # b2 的估計值
se_b2 <- 0.00215        # b2 的標準誤

# 計算 t-統計量
t_value <- (b2_estimate - 0) / se_b2

# 設定自由度
df <- 62

# 查找 1% 顯著性水平下的單尾 t 分布臨界值
critical_value <- qt(0.99, df)

# 輸出 t-值和臨界值
print(paste("t-value:", t_value))
print(paste("Critical value at 1% significance level:", critical_value))

# 判斷是否拒絕零假設
if (t_value > critical_value) {
  print("Reject the null hypothesis: There is a significant positive relationship between GDP and the number of medals.")
} else {
  print("Do not reject the null hypothesis: There is no significant positive relationship between GDP and the number of medals.")
}
```

```
## [1] "t-value: 6.08837209302326"
## [1] "Critical value at 1% significance level: 2.38801077482455"
## [1] "Reject the null hypothesis: There is a significant positive relationship between GDP and the number of medals."
```

# Q1 (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?

## Ans

- By t-test $t = 6.088 > 2.388$ we reject the null hypothesis ].
- my conclusion is that higher GDP significantly contributes to winning more Olympic medals.
- A 1% level of significance means there’s only a 1% chance of wrongly rejecting a true null hypothesis (Type I error). This high confidence level shows that we are 99% sure of our findings that GDP has a significant positive effect on medal counts.

# Q7

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

$$I\hat{NCOME} = (a) + 1.029BACHELOR$$

# Q7 (a)

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

# Ans

$a = t * se(a) = 4.31 * 2.672 = 11.51632$

# Q7 (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?

# Ans

- $I\hat{NCOME} = 11.51632 + 1.029BACHELOR$
- Increasing, since slope $1.029 > 0$.
- positive relationship, also the slope > 0.
- it is increasing at a constant rate.

# Q7 (c)

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

# Ans

$t = \frac{\text{slope}-\text{hypothesis value}}{c} = \frac{\text{slope}}{c}\ c = \frac{\text{slope}}{t} = \frac{1.029}{10.75} = 0.09572$

# Q7 (d)

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

# Ans

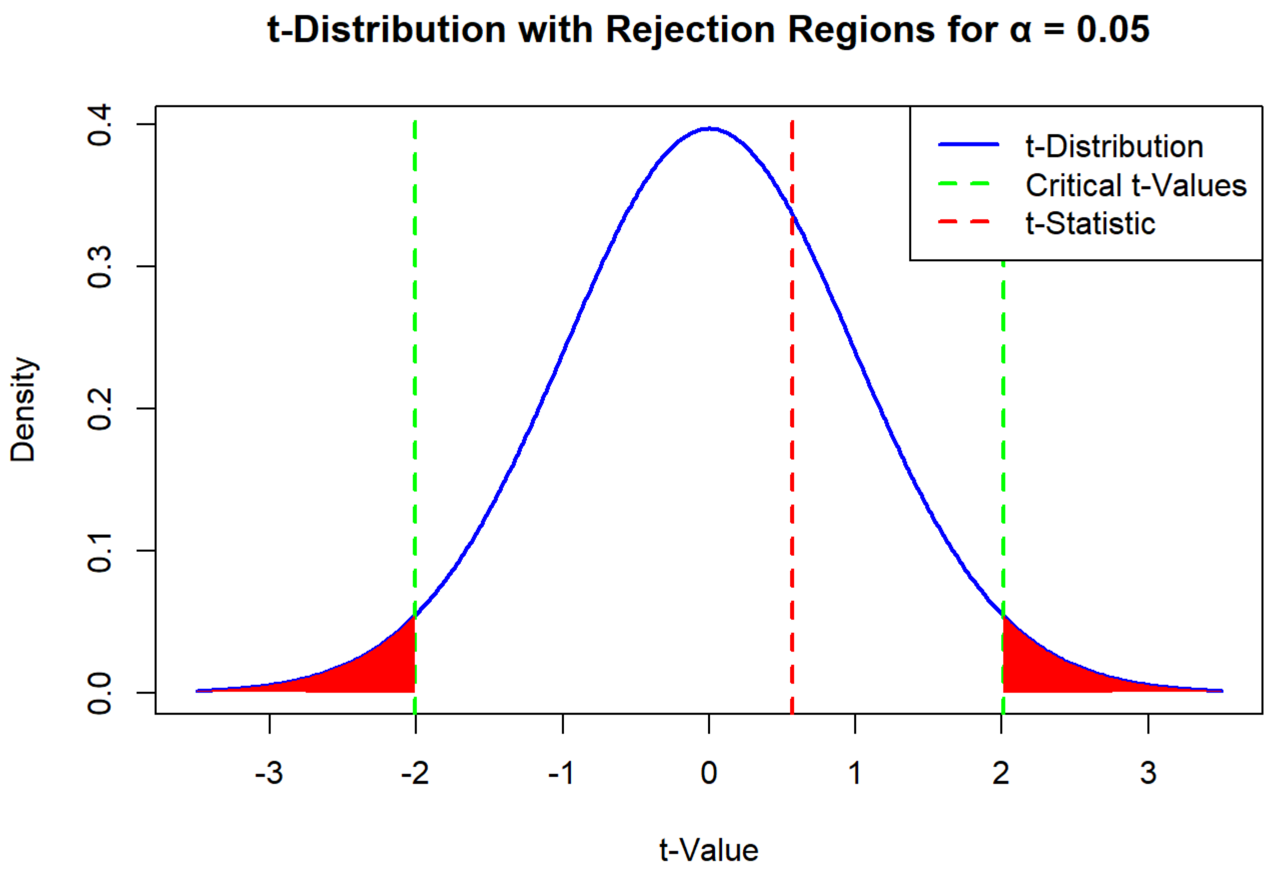
$t = \frac{\text{estimate}-\text{hypothesis value}}{se} = \frac{11.51632-10}{2.672} = 0.5675$

# Q7 (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 $\alpha = 0.05$.

# Ans

```
# 設定參數
t_statistic <- 0.567
alpha <- 0.05
df <- 50

# 計算臨界 t 值
critical_t_positive <- qt(1 - alpha / 2, df)
critical_t_negative <- qt(alpha / 2, df)

# 生成 t 分佈的 x 和 y 值
x <- seq(-3.5, 3.5, length.out = 400)
y <- dt(x, df)

# 繪圖設定
plot(x, y, type = 'l', col = 'blue', lwd = 2,
     main = 't-Distribution with Rejection Regions for α = 0.05',
     xlab = 't-Value', ylab = 'Density')

# 畫出臨界 t 值（拒絕區域的邊界）
abline(v = critical_t_positive, col = 'green', lty = 'dashed', lwd = 2)
abline(v = critical_t_negative, col = 'green', lty = 'dashed', lwd = 2)

# 標示 t-統計值
abline(v = t_statistic, col = 'red', lty = 'dashed', lwd = 2)

# 標示拒絕區域
x_fill_neg <- c(x[x <= critical_t_negative], critical_t_negative)
y_fill_neg <- c(y[x <= critical_t_negative], 0)
polygon(x_fill_neg, y_fill_neg, col = 'red', border = NA)

x_fill_pos <- c(critical_t_positive, x[x >= critical_t_positive])
y_fill_pos <- c(0, y[x >= critical_t_positive])
polygon(x_fill_pos, y_fill_pos, col = 'red', border = NA)

# 添加圖例
legend("topright", legend = c("t-Distribution", "Critical t-Values", "t-Statistic"),
       col = c("blue", "green", "red"), lty = c(1, 2, 2), lwd = 2)
```



# Q7 (f)

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.

## Ans

$t = \frac{1.029-1}{0.0957} = 0.303$

The $t$-statistic calculated is $0.303$ with a $p$-value of $0.763$. Since the $p$-value is much higher than the significance level of $0.05$, we do not reject the null hypothesis that the slope coefficient is $1$. This suggests that the effect of the percentage of the population with a bachelor’s degree on per capita income is statistically not different from $\$1000$ per $1$ increase in bachelor’s degree holders.

# Q17

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 $W\hat{A}GE = -10.76 + 2.46 EDUC, N = 986$ (se) = (2.27) (0.16) Rural $W\hat{A}GE = -4.88 + 1.80 EDUC, N = 214$ (se) = (2.27) (0.16)

# Q17 (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 α = 0.05 level of significance. Show all steps, including a graph of the critical region and state your conclusion.

# Ans

1.

$$H_0 : \beta_2 = 1.80$$

$$H_\alpha : \beta_2 > 1.80$$

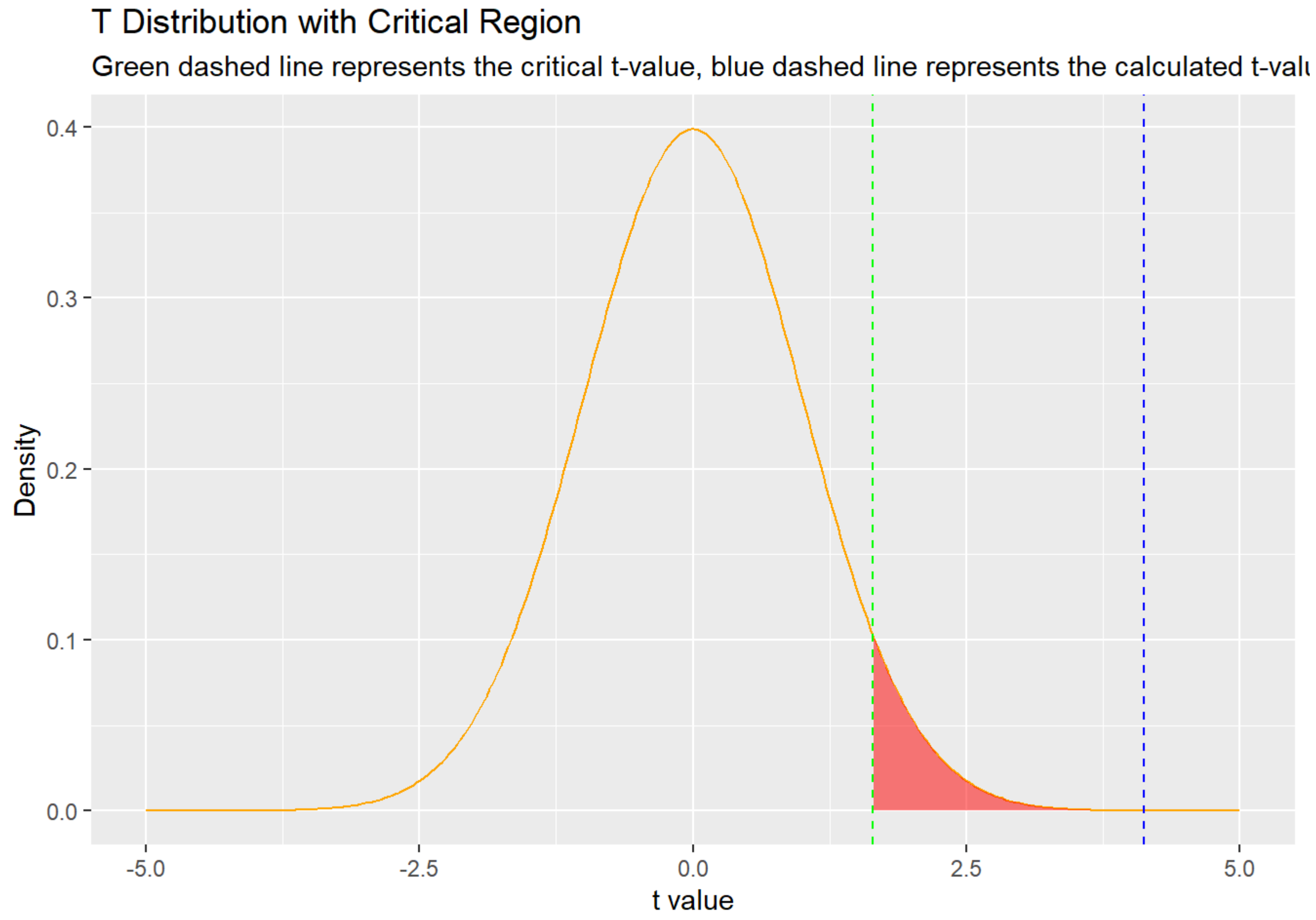
2.

$$t = \frac{2.46 - 1.80}{0.16} = 4.125$$

3. At a significance level of $\alpha = 0.05$, for a $t$-distribution with degrees of freedom $N - 2 = 984$, the critical value obtained from the $t$-table is approximately $1.645$. This is a one-tailed test.

4.

```r
# 設定參數
df <- 984  # 自由度
alpha <- 0.05  # 顯著性水平
t_value <- 4.125  # 計算出的 t 值
t_critical <- qt(1 - alpha, df)  # 臨界值

# 生成 t 分佈數據
t_values <- seq(-5, 5, length.out = 300)  # t 值範圍
t_density <- dt(t_values, df)  # 計算 t 分佈的密度函數

# 建立資料框
t_df <- data.frame(t_values, t_density)

# 創建圖表
ggplot(t_df, aes(x = t_values, y = t_density)) +
  geom_line(color = "orange") +  # 繪製 t 分佈曲線
  geom_area(data = subset(t_df, t_values > t_critical), aes(x = t_values, y = t_density), fill = "red", alpha = 0.5) +  # 標示臨界區域
  geom_vline(xintercept = t_critical, color = "green", linetype = "dashed") +  # 標示臨界 t 值
  geom_vline(xintercept = t_value, color = "blue", linetype = "dashed") +  # 標示計算的 t 值
  labs(title = "T Distribution with Critical Region",
       subtitle = "Green dashed line represents the critical t-value, blue dashed line represents the calculated t-value") +
  xlab("t value") +
  ylab("Density")
```


T Distribution with Critical Region

Green dashed line represents the critical t-value, blue dashed line represents the calculated t-valu

5. Since the calculated t-value of 4.125 is much greater than the critical value of 1.645, we reject the null hypothesis at the 0.05 significance level. There is sufficient evidence to conclude that the slope is greater than 1.80. This implies that the positive impact of education on hourly wages is more significant in urban areas compared to rural areas.

# Q17 (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.

## Ans

1. $\hat{WAGE} = -4.88 + 1.80EDUC = -4.88 + 1.80 * 16 = 23.92$
2. find $t_{\alpha/2,214-2}$. which is $1.971$
3. $CI = \hat{WAGE} \pm t_{\alpha/2} * se(\hat{WAGE}) = 23.92 \pm 1.971 * 0.833$
4. $(22.278, 25.562)$

# Q17 (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.

## Ans

1. $\hat{WAGE} = -10.76 + 2.46 \times EDUC = -10.76 + 2.46 \times 16 = -10.76 + 39.36 = 28.60$

2. $SE(W\hat{AGE}) = \sqrt{Var(\hat{\beta}_1) + x^2Var(\hat{\beta}_2) + 2x \cdot Cov(\hat{\beta}_1, \hat{\beta}_2)} = 0.8164$
3. $t_{0.05/2,984} = 1.961$
4. CI of urban is $(26.998, 30.202)$

- the interval estimate for the urban regression narrower than that for the rural regression in (b)
- since N of urban is larger $986 > 214$. the esitmation is more accurate.

# Q17 (d)

Using the rural regression, test the hypothesis that the intercept parameter β1 equals four, or more, against the alternative that it is less than four, at the 1% level of significance. ### Ans 1. - $H_0 : \beta_1 \geq 4$ - $H_1 : \beta_1 < 4$ 2. $t = \frac{\text{esitmate}-\text{assumption}}{se} = \frac{-4.88-4}{3.29} = -2.699 < 4$ 3. find $t_{\alpha,214-2} = t_{0.01,212} = -2.344$ 4. since $-2.699 < -2.344$, we reject $H_0$

At the **1% significance level**, we have sufficient evidence to conclude that the intercept parameter $\beta_1$ is less than 4. This suggests that when **EDUC = 0** (no education), the predicted **WAGE** is reasonably assumed to be less than **$4 per hour**.

# Q19

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.

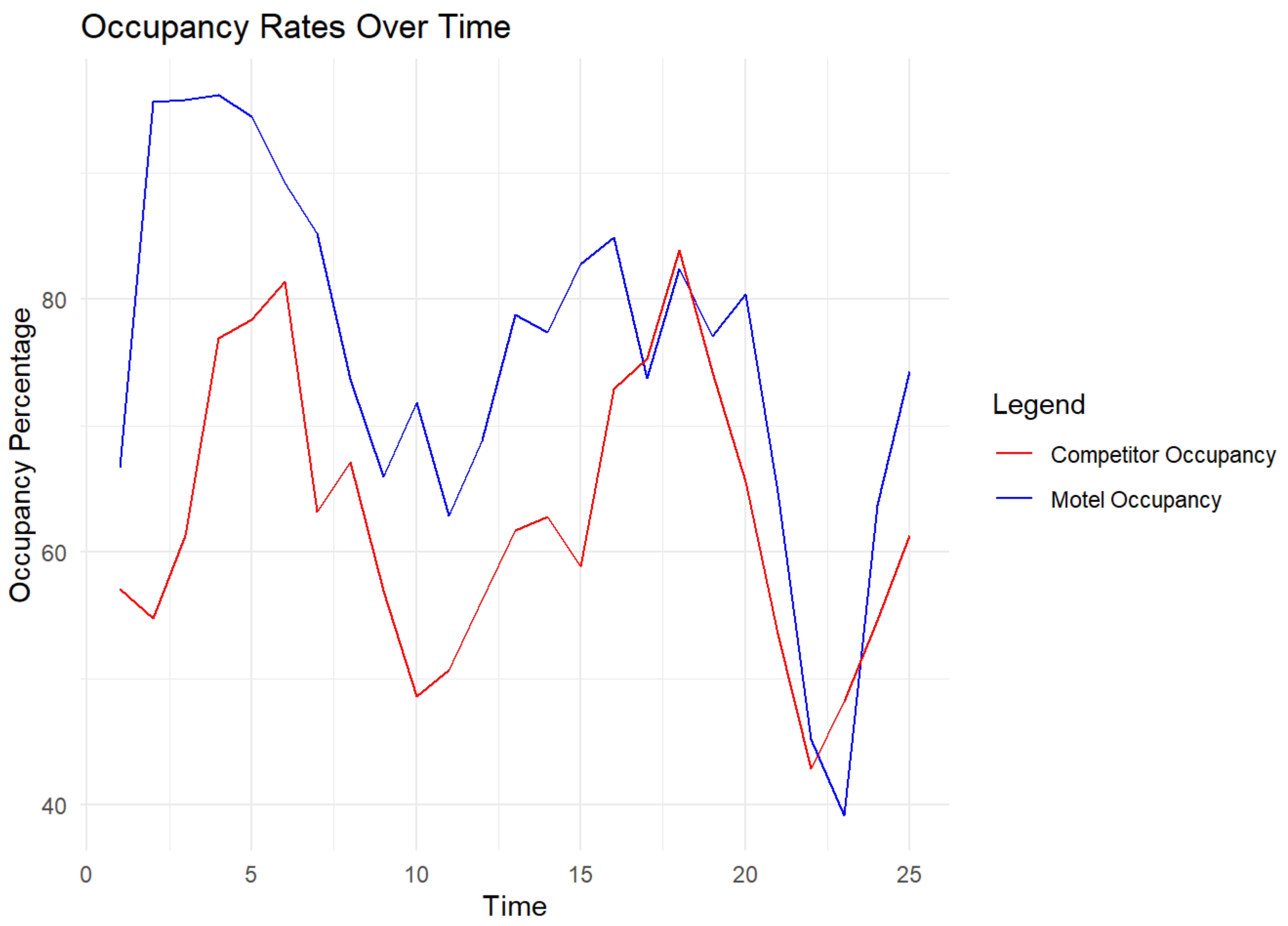
```
data(motel)
```

# Q19 (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 = β1 + β2COMP_PCT + e. Construct a 95% interval estimate for the parameter β2. Have we estimated the association between MOTEL_PCT and COMP_PCT relatively precisely, or not? Explain your reasoning.

## Ans

```
# 加載必要的套件
library(ggplot2)

# 1. 繪製MOTEL_PCT 和 COMP_PCT 隨時間變化的趨勢圖
ggplot(motel, aes(x = time)) +
  geom_line(aes(y = motel_pct, color = "Motel Occupancy")) +
  geom_line(aes(y = comp_pct, color = "Competitor Occupancy")) +
  labs(title = "Occupancy Rates Over Time",
       x = "Time",
       y = "Occupancy Percentage") +
  scale_color_manual(values = c("Motel Occupancy" = "blue", "Competitor Occupancy" = "red"),
                     name = "Legend") +
  theme_minimal()
```



```
# 2. 執行線性回歸
model <- lm(motel_pct ~ comp_pct, data = motel)

# 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
```

```
# 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
```

- What can you say about the occupancy rates over time?
  - the occupancy rates over time go up and down.
- Do they tend to move together?
  - yes, they do.
- Which seems to have the higher occupancy rates?
  - MOTEL_PCT
- Construct a 95% interval estimate for the parameter β2.
  - (0.4452978 1.283981)
- Have we estimated the association between MOTEL_PCT and COMP_PCT relatively precisely, or not? Explain your reasoning.
  - no , since CI is width.

# Q19 (b)

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

## Ans

```
# 設定 comp_pct = 70
new_data <- data.frame(comp_pct = 70)

# 預測 MOTEL_PCT 並建立 90% 信賴區間
pred <- predict(model, newdata = new_data, interval = "confidence", level = 0.90)

# 顯示結果
pred
```

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

# Q19 (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.

# Ans

```r
# 提取 β2 的估計值、標準誤和 t 值
beta2_hat <- coef(summary(model))["comp_pct", "Estimate"]
se_beta2 <- coef(summary(model))["comp_pct", "Std. Error"]
t_value <- coef(summary(model))["comp_pct", "t value"]

# 設定自由度 df = n - 2
df <- nrow(motel) - 2

# 計算臨界值 (alpha = 0.01，右尾檢定)
t_critical <- qt(0.99, df)

# 顯示 t 值與臨界值
cat("t-value:", t_value, "\n")
cat("Critical t-value at alpha = 0.01:", t_critical, "\n")

# 判斷是否拒絕 H0
if (t_value > t_critical) {
  cat("結論: 拒絕 H0，COMP_PCT 對 MOTEL_PCT 具有顯著的正向影響。\n")
} else {
  cat("結論: 無法拒絕 H0，無足夠證據證明 COMP_PCT 對 MOTEL_PCT 具有正向影響。\n")
}
```

```
## t-value: 4.26536
## Critical t-value at alpha = 0.01: 2.499867
## 結論: 拒絕 H0，COMP_PCT 對 MOTEL_PCT 具有顯著的正向影響。
```

# Q19 (d)

In the linear regression model $MOTEL_{P}CT = \beta_1 + \beta_2 COMP_{P}CT + e$, test the null hypothesis $H_0 : \beta_2 = 1$ against the alternative hypothesis \$H_0:β_2 ≠ \$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.

## Ans

```
# 設定假設 H0: β2 = 1
beta2_0 <- 1

# 計算 t 統計量
t_value <- (beta2_hat - beta2_0) / se_beta2

# 設定自由度 df = n - 2
df <- nrow(motel) - 2

# 計算雙尾檢定的臨界值 (alpha = 0.01)
t_critical <- qt(0.995, df)  # 雙尾檢定，alpha/2 = 0.005

# 顯示 t 值與臨界值
cat("t-value:", t_value, "\n")
cat("Critical t-value at alpha = 0.01:", c(-t_critical, t_critical), "\n")

# 判斷是否拒絕 H0
if (abs(t_value) > t_critical) {
  cat("結論: 拒絕 H0，β2 顯著不同於 1。\n")
} else {
  cat("結論: 無法拒絕 H0，沒有足夠證據表明 β2 與 1 顯著不同。\n")
}
```

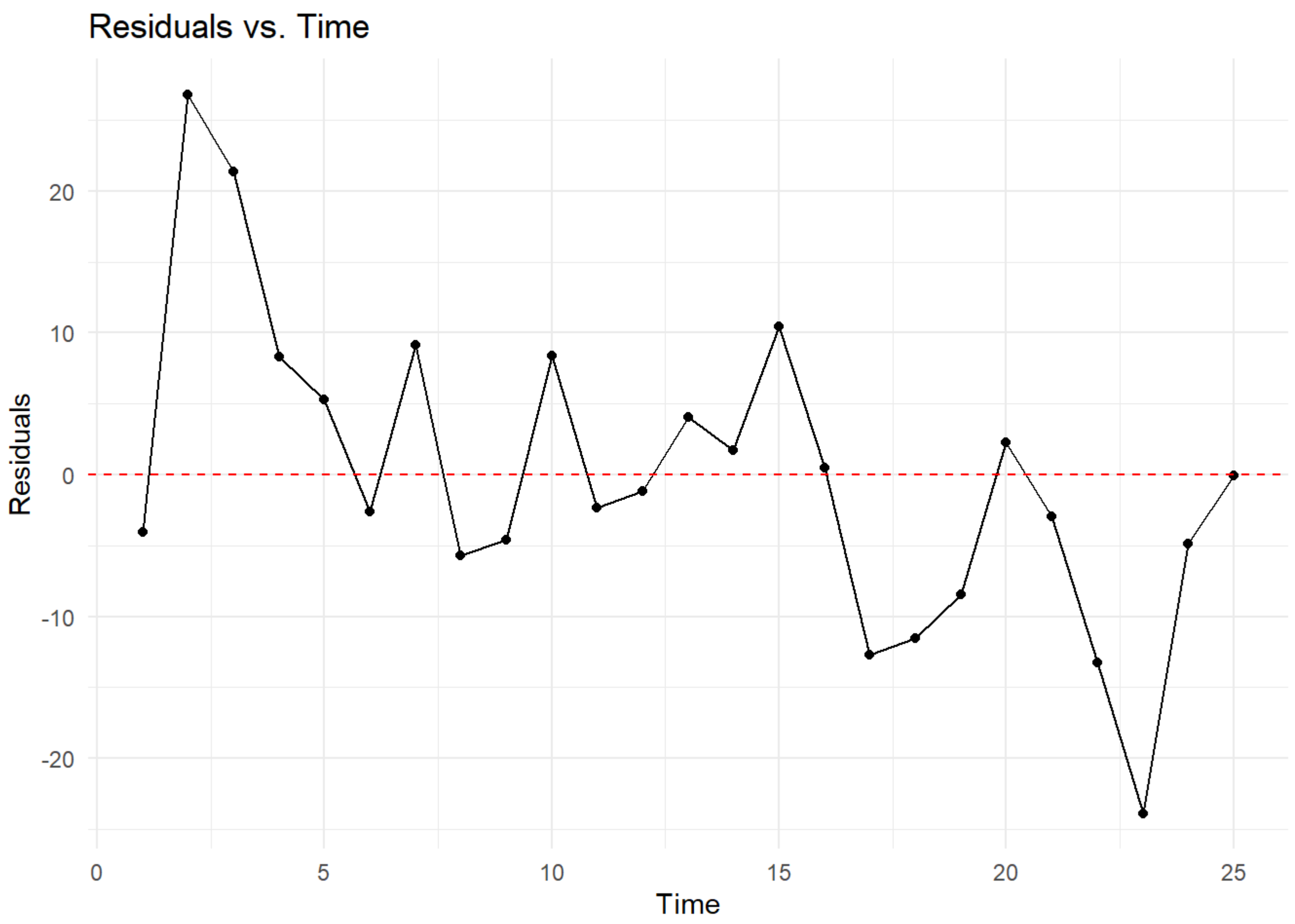
```
## t-value: -0.6677491
## Critical t-value at alpha = 0.01: -2.807336 2.807336
## 結論: 無法拒絕 H0，沒有足夠證據表明 β2 與 1 顯著不同。
```

# Q19 (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)?

## Ans

```
# 計算殘差
motel$residuals <- residuals(model)

# 繪製殘差對時間的圖
ggplot(motel, aes(x = time, y = residuals)) +
  geom_point() +
  geom_line() +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residuals vs. Time",
       x = "Time",
       y = "Residuals") +
  theme_minimal()
```



```
# 檢查 17-23 期間的殘差
subset_residuals <- motel$residuals[motel$time %in% 17:23]
subset_residuals
```

```
## [1] -12.707328 -11.543226  -8.456225   2.279673  -2.958191 -13.293015 -23.875603
```

```
mean(subset_residuals > 0)  # 計算正殘差比例
```

```
## [1] 0.1428571
```

- **Anomalous Features**:
  - **During Time 17-23, residuals are significantly negative, indicating that `MOTEL_PCT` is lower than the model's predicted values**.
  - **During Time 1-10, most residuals are positive, suggesting that `MOTEL_PCT` is higher than the model's predicted values**.
  - **The most negative residuals occur during Time 20-23, indicating that the impact of construction may be at its peak**.
- **85.71% of the residuals during Time 17-23 are negative**, suggesting that the model **significantly overestimated** `MOTEL_PCT` during this period. This may be due to the impact of construction, which reduced occupancy rates, causing the actual values to be much lower than the predicted values.