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

(b) 檢定統計量: $T = \dfrac{b_2 - 0}{\sqrt{\frac{MSE}{SXX}}} = \dfrac{b_2}{SE(b_2)} \overset{H_0}{\sim} t(64-2)$

(C) don't reject H_0 ⇒ expect value of $b_2 = 0$

Accept H_1 ⇒ expect value of $b_2 > 0$

觀察到的統計量可能大於0

∴ t分佈的中心會向右偏移

(d) $\alpha = 0.01$

reject region: $\{T > t_{0.01}(62) = 2.388\}$

$T_0 \in RR$, 則 reject H_0

$T_0 \notin RR$, 則 don't reject H_0

(e) $T_0 = \frac{b_2}{SE(b_2)} = \frac{0.01309}{0.00215} \approx 6.084$

$\because T_0 = 6.084 > 2.388$

$\therefore$ accept H_1, there is a positive relationship between the number of medals won and GDP

$0.01 = \alpha = p(\text{Type I error}) = p(\text{reject } H_0 \mid H_0 \text{ is true})$

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

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

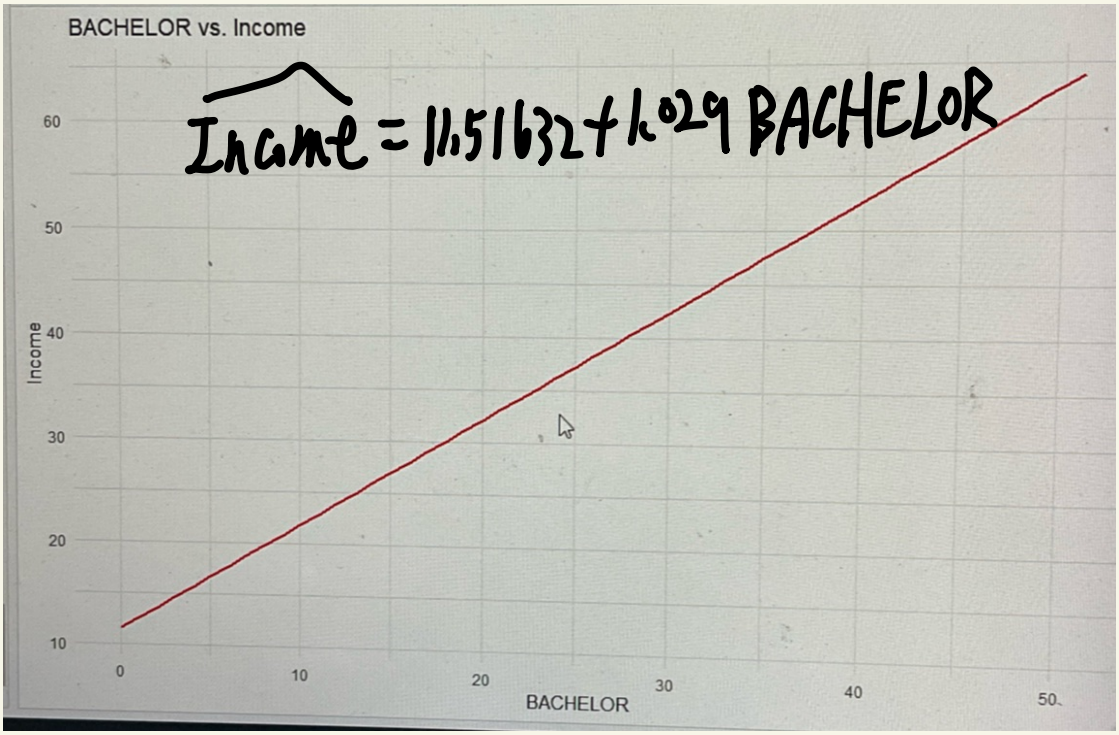
a. Using the information provided calculate the estimated intercept. Show your work.

(a) $t = \frac{b_1}{se(b_1)} \Rightarrow b_1 = t \cdot se(b_1) = 2.672 \times 4.31$
$= 11.51632$

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?

$\because \beta_2 > 0 \therefore$ is increasing and is positive relationship

$\because$ it is a linear relationship (a straight line) $\therefore$ it is increasing at a constant rate.



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

$$t = \frac{b_2}{se(b_2)} \Rightarrow se(b_2) = \frac{b_2}{t} = \frac{1.029}{10.75} = 0.0957$$

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

$$H_0: \beta_1 = 10 \quad H_1: \beta_1 \neq 10$$

$$T_0 = \frac{b_1 - 10}{se(b_1)} = \frac{11.5632 - 10}{2.672} = 0.587$$

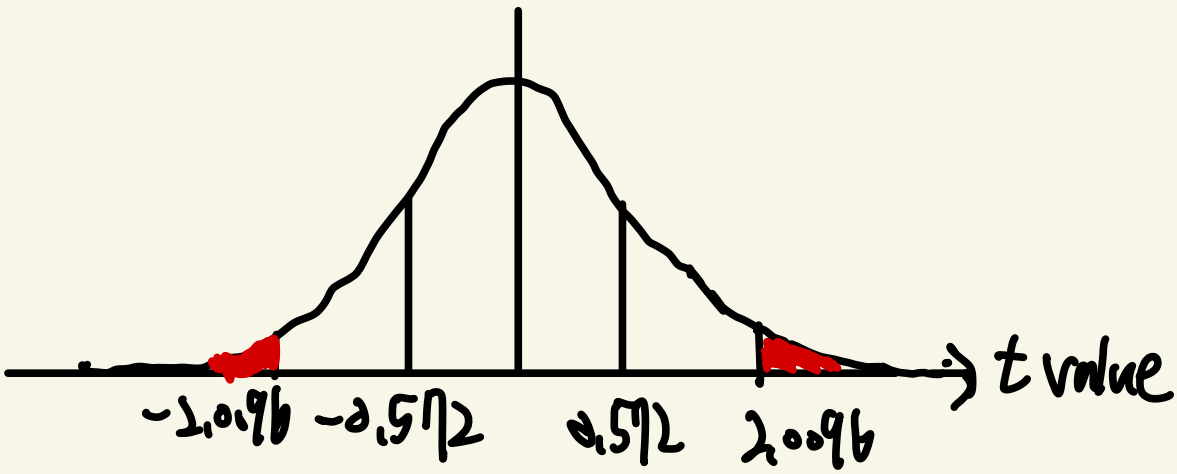
e. The *p*-value for a two-tail test that the intercept parameter equals 10, from part (d), is 0.572. Show the *p*-value in a sketch. On the sketch, show the rejection region if $\alpha = 0.05$.

$$\alpha = 0.05$$

$$\text{Reject region} = \{|T| \geq t_{0.05}(57-2)\}$$

$$= \{T \geq t_{0.05}(49) \text{ or } T \leq -t_{0.05}(49)\}$$

$$= \{T \geq 2.096 \text{ or } T \leq -2.096\}$$



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

$$1-0.01=P\left(-t_{0.005}(49)<\frac{1.029-\beta_2}{0.0957}<t_{0.005}(49)\right)$$

$$=P\left(1.029-t_{0.005}(49)\cdot 0.0957<\beta_2<1.029+t_{0.005}(49)\cdot 0.0957\right)$$

β_2的99% CI $=[1.029-2.6449\cdot 0.0957,\ 1.029+2.645\cdot 0.0957]$

$=[0.7725,\ 1.2855]$ $(t_{0.05}(49)=2.6449)$

我們有99%的信心，區間$[0.7725, 1.2855]$會包含真實的母體斜率β_2

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.

$H_0:\beta_2=1 \quad H_1:\beta_2\neq 1$

$\alpha=0.05$

檢定統計量：$T=\frac{b_2-1}{SE(b_2)}$ 此 $t_{0.05}(49)$

$RR=\{T\geq 2.0096 \text{ or } T\leq -2.0096\}$

$T_0=\frac{1.029-1}{0.0957}=0.303$

$\because T_0$ & RR $\therefore$ Don't reject H_0

There is no evidence to say the slope coefficient (β_2) is not one.

observation in part (d).

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

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

$H_0: \beta_2 = 1.80$

$H_1: \beta_2 > 1.80$

$\alpha = 0.05$

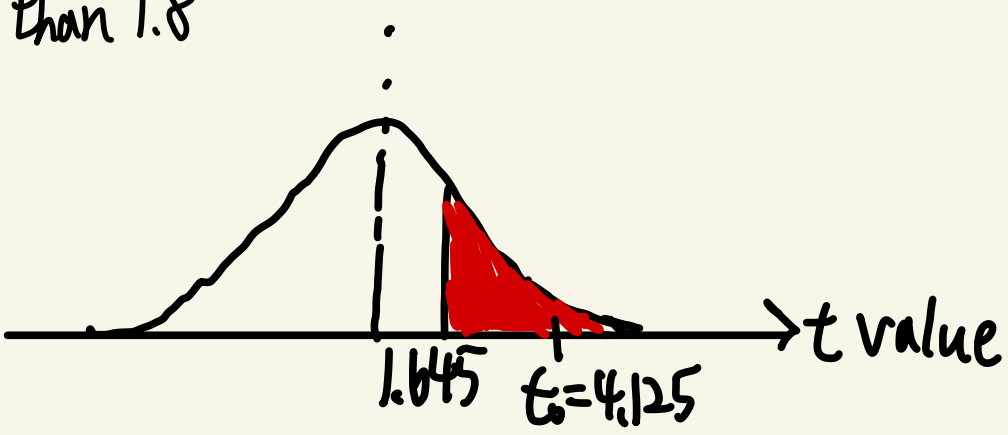
檢定統計量: $T = \frac{b_2 - 1.8}{SE(b_2)} \overset{H_0}{\sim} t_{0.05}(986-2)$

$RR = \{T > t_{0.05}(984) = 1.65\}$

$T_0 = \frac{2.46 - 1.8}{0.16} = 4.125$

$\because T_0 \in RR \quad \therefore$ Reject H_0

There is an evidence to say that β_2 is 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.

$\hat{y} = \widehat{wage} = -4.88 + 1.8 \cdot 16 = 23.92$, let $x = EDUC$

檢定統計量 $T = \dfrac{\hat{y} - (\beta_1 + \beta_2 x)}{se(\hat{y})} \sim t_{(214-2)}$

$0.95 = P\left(-t_{0.025(212)} < \dfrac{23.92 - (\beta_1 + \beta_2 x)}{0.8333} < t_{0.025(212)}\right)$

$= P(23.92 - t_{0.025(212)} \cdot 0.8333 < \beta_1 + \beta_2 x < 23.92 + t_{0.025(212)} \cdot 0.8333)$

$t_{0.025(212)} \simeq 1.97$

$CI = [23.92 - 1.97 \cdot 0.8333, 23.92 + 1.97 \cdot 0.8333]$

$= [22.28, 25.56]$

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.

做法同(b)

$\hat{y} = -10.76 + 2.46 \cdot 16 = 28.6$

$se(\hat{y}) = \sqrt{se^2_{\beta_1} + (EDUC)^2 se^2_{\beta_2} + 2(EDUC) \cdot cov(\beta_1, \beta_2)}$

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

$= \sqrt{0.6665} \simeq 0.8164$

$t_{0.025}(984) \approx 1.96$

$\therefore CI = [\hat{Y} - t_{0.025}(984) \cdot SE(\hat{Y}),\ \hat{Y} + t_{0.025}(984) \cdot SE(\hat{Y})]$

$= [28.6 - 1.96 \cdot 0.8164,\ 28.6 + 1.96 \cdot 0.8164]$

$= [27.00, 30.20]$

the CI for the urban regression is narrower.

∵ standard error is smaller than rural regression ($0.8164 < 0.8333$)

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.

$H_0: \beta_1 \geq 4$

$H_1: \beta_1 < 4$

$\alpha = 0.01$

檢定統計量: $T = \frac{b_1 - 4}{se(b_1)} \overset{H_0}{\sim} t_{0.01}(212)$

$RR = \{T < t_{0.01}(212) \approx -2.34\}$

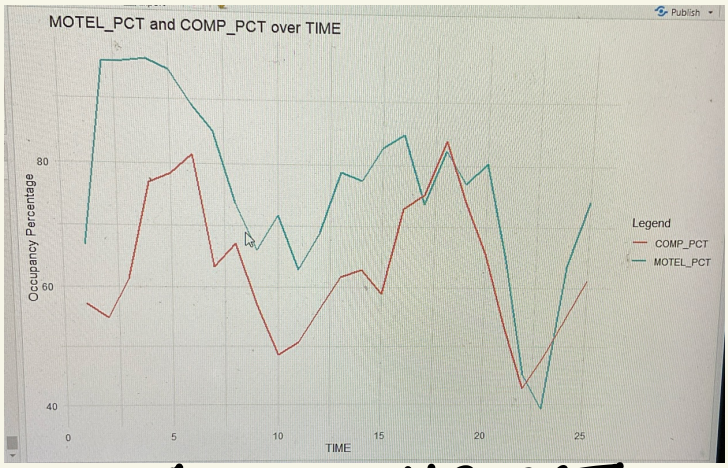
$T_0 = \frac{-4.88 - 4}{3.29} = -2.7$

∵ $T_0 \in RR$ ∴ Reject H_0

There is an evidence to say that β_1 is less than four.

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_PCT and COMP_PCT tend to move together and MOTEL_PCT seems to have higher occupancy rates

```
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

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

MOTEL_PCT = 21.4 + 0.865 COMP_PCT

95% CI of β_2 = [0.4453, 1.284]

∵ $\beta_2 \approx 1$ 代表兩者變動趨勢幾乎一致

∴ 估 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.

```
> predict(model, new_data, interval = "confi
       fit      lwr      upr
1 81.92474 77.38223 86.46725
>
```

90% CI = [77.38, 86.47]

c. In the linear regression model $MOTEL\_PCT = \beta_1 + \beta_2 COMP\_PCT + e$, test the null hypothesis $H_0: \beta_2 \leq 0$ 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.

檢定統計量 $T = \frac{b_2 - 0}{se(b_2)} \overset{H_0}{\sim} t_{0.01}(23)$

```
> # 提取回歸係數、標準誤差和 p 值
> beta_2 <- coef(summary(model))["comp_pct", "Estimate"]
> se_beta_2 <- coef(summary(model))["comp_pct", "Std. Error"]
> t_value <- beta_2 / se_beta_2  # 計算 t 統計量
> # 計算 p 值（單尾檢定）
> p_value <- pt(t_value, df = model$df.residual, lower.tail = FALSE)
> # 顯示結果
> cat("t 統計量:", t_value, "\n")
t 統計量: 4.26536
> cat("p 值:", p_value, "\n")
p 值: 0.0001453107
> # 判斷是否拒絕 H0
> if (p_value < 0.01) {
+   cat("結論: 拒絕 H0，β2 顯著大於 0。\n")
+ } else {
+   cat("結論: 無法拒絕 H0，β2 不顯著大於 0。\n")
+ }
結論: 拒絕 H0，β2 顯著大於 0。
> View(model)
> View(model)
```

$RR = \{T > t_{0.01}(23) = 2.5\}$

$T_0 = 4.265$

∵ p-value = 0.00014 < α = 0.01

∴ reject H_0, There is a evidence to say $\beta_2 > 0$

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.

檢定統計量: $T = \frac{b_2 - 1}{se(b_2)}$ 也 $t_{0.005}(23)$

```
> # 提取回歸係數與標準誤
> beta_2 <- coef(summary(model))["comp_pct", "Estimate"]
> se_beta_2 <- coef(summary(model))["comp_pct", "Std. Error"]
> # 計算 t 統計量
> t_value <- (beta_2 - 1) / se_beta_2
> # 計算 p 值（雙尾檢定）
> p_value <- 2 * pt(abs(t_value), df = model$df.residual, lower.tail = FALSE)
> # 顯示結果
> cat("t 統計量:", t_value, "\n")
t 統計量: -0.6677491
> cat("p 值:", p_value, "\n")
p 值: 0.5109392
> # 判斷是否拒絕 H0
> if (p_value < 0.01) {
+   cat("結論: 拒絕 H0，β2 顯著不同於 1。\n")
+ } else {
+   cat("結論: 無法拒絕 H0，β2 不顯著不同於 1。\n")
+ }
結論: 無法拒絕 H0，β2 不顯著不同於 1。
>
```

$RR = \{|T| \geq t_{0.05}(23)\} = 2.807\}$

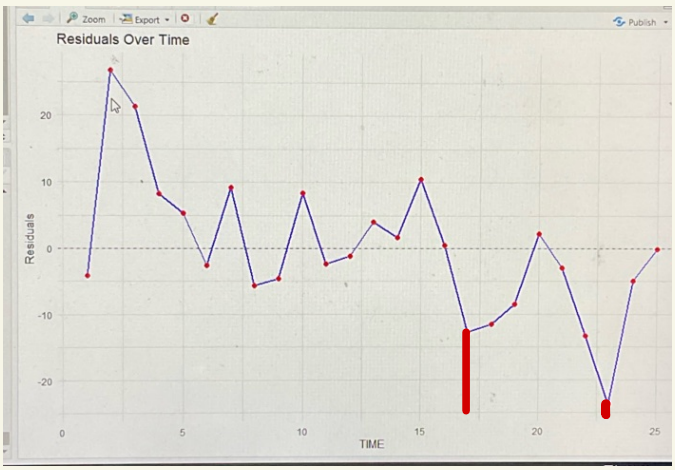
$T_0 = -0.67$

$\because p\text{-value} = 0.511 > \alpha = 0.01$

$\therefore$ Don't reject H_0, there is no evidence to say $\beta_2 \neq 1$

imply that model's occupancy rate is similar to competitor's occupancy rate.

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)?



All the residuals are negative expect for 23 during time period 17-23

3.19 R

```r
#3.19a
ggplot(motel, aes(x = time)) +
  geom_line(aes(y = motel_pct, color = "MOTEL_PCT"), size = 1) +  # MOTEL_PCT 折線
  geom_line(aes(y = comp_pct, color = "COMP_PCT"), size = 1) +    # COMP_PCT 折線
  labs(title = "MOTEL_PCT and COMP_PCT over TIME",
       x = "TIME",
       y = "Occupancy Percentage",
       color = "Legend") +
  theme_minimal()
model <- lm(motel_pct ~ comp_pct, data = motel)
summary(model)
confint(model, level = 0.95)
#b
new_data <- data.frame(comp_pct = 70)
predict(model, new_data, interval = "confidence", level = 0.90)
#c
summary(model)

# 提取回歸係數、標準誤差和 p 值
beta_2 <- coef(summary(model))["comp_pct", "Estimate"]
se_beta_2 <- coef(summary(model))["comp_pct", "Std. Error"]
t_value <- beta_2 / se_beta_2  # 計算 t 統計量

# 計算 p 值（單尾檢定）
p_value <- pt(t_value, df = model$df.residual, lower.tail = FALSE)

# 顯示結果
cat("t 統計量:", t_value, "\n")
cat("p 值:", p_value, "\n")

# 判斷是否拒絕 H0
if (p_value < 0.01) {
  cat("結論: 拒絕 H0，β2 顯著大於 0。\n")
} else {
  cat("結論: 無法拒絕 H0，β2 不顯著大於 0。\n")
}

#d
# 提取回歸係數與標準誤
beta_2 <- coef(summary(model))["comp_pct", "Estimate"]
se_beta_2 <- coef(summary(model))["comp_pct", "Std. Error"]

# 計算 t 統計量
t_value <- (beta_2 - 1) / se_beta_2

# 計算 p 值（雙尾檢定）
p_value <- 2 * pt(abs(t_value), df = model$df.residual, lower.tail = FALSE)

# 顯示結果
cat("t 統計量:", t_value, "\n")
cat("p 值:", p_value, "\n")

# 判斷是否拒絕 H0
if (p_value < 0.01) {
  cat("結論: 拒絕 H0，β2 顯著不同於 1。\n")
} else {
  cat("結論: 無法拒絕 H0，β2 不顯著不同於 1。\n")
}

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

# 繪製殘差隨時間變化的折線圖
ggplot(motel, aes(x = time, y = residuals)) +
  geom_line(color = "blue", size = 1) +
  geom_point(color = "red", size = 2) +  # 標示各點
  geom_hline(yintercept = 0, linetype = "dashed", color = "black") + # 加入零軸
  labs(title = "Residuals Over Time",
       x = "TIME",
       y = "Residuals") +
  theme_minimal()
```