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?

(SOLN)

a. $\begin{cases} H_0: b_2 = 0 \\ H_a: b_2 > 0 \end{cases}$

b. $n = 64$

test statistic: $T = \dfrac{\hat{\beta}_2}{\sqrt{\dfrac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \overset{H_0}{\sim} t(64-2)$

c. $RR = \{T > t_\alpha(62)\}$

If H_a is true, $T_0 \in RR$, the expected value of $b_2 > 0$.

the distribution will shift to right.

d. $\alpha = 0.01 \Rightarrow t_{0.01}(62) = 2.388$

We will reject H_0 if $T \geq 2.388$; do not reject H_0 if $T < 2.388$

e. $T_0 = \dfrac{\hat{\beta}_2}{SE(\hat{\beta}_2)} = \dfrac{0.01309}{0.00215} \approx 6.088 \in RR$, reject H_0

統計上, the number of medals won and GDP 存在顯著正相關

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

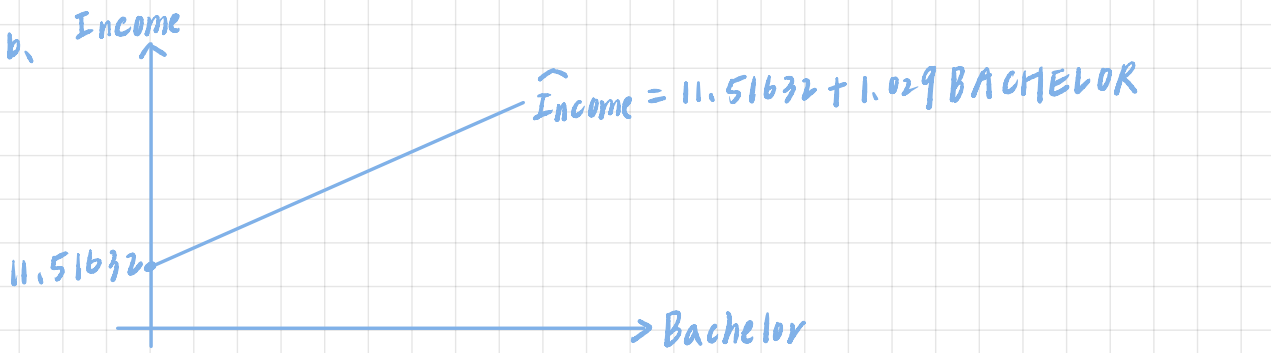
$$\begin{array}{lccc} \widehat{INCOME} = & (a) & + & 1.029BACHELOR \\ \text{se} & (2.672) & & (c) \\ t & (4.31) & & (10.75) \end{array}$$

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

(SOLN)

a. $t = \dfrac{\hat{\beta}_0}{SE(\hat{\beta}_0)} \Rightarrow 4.31 = \dfrac{\hat{\beta}_0}{2.672}$ $\therefore \hat{\beta}_0 = 11.51632$

b.



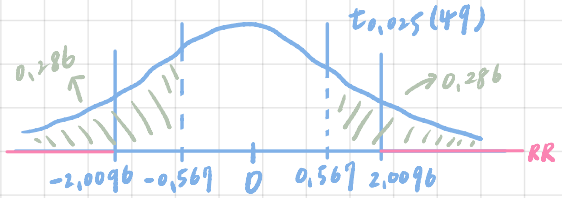
Income and bachelor have a positive relationship and increasing at constant rate because the b_2 coefficient is 1.029

c. $t = \dfrac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \Rightarrow 10.75 = \dfrac{1.029}{SE(\hat{\beta}_1)}$ $\therefore SE(\hat{\beta}_1) = 0.0957$

d. $\begin{cases} H_0: \beta_0 = 10 \\ H_a: \beta_0 \neq 10 \end{cases}$

$$T = \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} \overset{H_0}{=} \frac{11.51632 - 10}{2.672} = 0.567$$

e. $RR = \{|T| \geq t_{0.025(51-2)} = t_{0.025(49)} = 2.0096\}$



f. interval estimate of the slope: $\hat{\beta}_1 \pm t_{\frac{\alpha}{2}}(n-2) \cdot SE(\hat{\beta}_1)$

$\hat{\beta}_1 \pm t_{0.005}(49) SE(\hat{\beta}_1)$

$\Rightarrow [1.029 - 2.680 \cdot 0.0957,\ 1.029 + 2.680 \cdot 0.0957]$

$\Rightarrow [0.772524,\ 1.285476]$

g. $\begin{cases} H_0: \beta_1 = 1 \\ H_a: \beta_1 \neq 1 \end{cases}$

$\alpha = 0.05$

t-statistic: $T = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \overset{H_0}{\sim} t(n-2)$

$RR = \{|T| \geq t_{\frac{\alpha}{2}}(n-2) = t_{0.025}(49) = 2.010\}$

$T_0 = \frac{1.029 - 1}{0.0957} = 0.303 \notin RR$, do not reject H_0

統計上, the slope 並未顯著異於 1,

即教育程度提升1%, 所得提升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.

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

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

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.

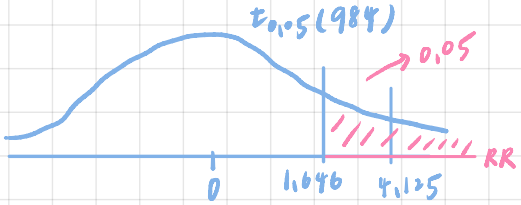
d. Using the rural regression, test the hypothesis that the intercept parameter β_1 equals four, or more, against the alternative that it is less than four, at the 1% level of significance.

(508)

a. $\begin{cases} H_0: \beta_1 = 1.8 \\ H_a: \beta_1 > 1.8 \end{cases}$

$\alpha = 0.05$

$T = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \overset{H_0}{\sim} t(n-2)$



$RR = \{T > t_{0.05}(986-2) = 1.646\}$

$T_0 = \frac{2.46 - 1.8}{0.16} = 4.125 \in RR$, reject H_0

統計上, the slope of Urban 顯著大於 1.8

b. $EDUC = 16 \Rightarrow WAGE = 23.92$

$t_{0.025}(214-2) = 1.971$

interval estimate of Rural:

$[23.92 - 1.971 \cdot 0.833, 2.392 + 1.971 \cdot 0.833] = [22.278, 25.562]$

$SE(WAGE) = \sqrt{[SE(\hat{\beta}_0)]^2 + (EDUC)^2[SE(\hat{\beta}_1)]^2 + 2 \cdot EDUC \cdot Cov(\beta_0, \beta_1)}$

$= 1.1035$

c. $EDUC = 16 \Rightarrow WAGE = 28.6$

$SE(WAGE) = \sqrt{[SE(\hat{\beta}_0)]^2 + (EDUC)^2[SE(\hat{\beta}_1)]^2 + 2 \cdot EDUC \cdot Cov(\beta_0, \beta_1)}$

$= 0.8164$

$t_{0.025}(986-2) = 1.962$

interval estimate of Urban:

$[28.6 - 0.8164 \cdot 1.962, 28.6 + 0.8164 \cdot 1.962] = [26.998, 30.202]$

Hence, Urban's interval estimate is narrower than rural, because urban have more sample.

d. $\begin{cases} H_0: \beta_1 = 4 \\ H_a: \beta_1 < 4 \end{cases}$

$\alpha = 0.01$

$T = \frac{\hat{\beta}_1 - 4}{SE(\hat{\beta}_1)} \overset{H_0}{\sim} t(214-2)$

$RR = \{T < t_{0.01}(212) = 2.345\}$

$T_0 = \frac{-4.88 - 4}{3.29} = -2.699 \in RR$, reject H_0

統計上, β_1 顯著異於 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.

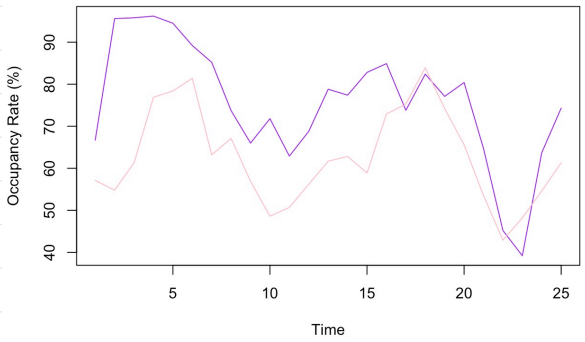
b. Construct a 90% interval estimate of the expected occupancy rate of the motel in question, *MOTEL_PCT*, given that $COMP\_PCT = 70$.

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

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.

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)?

(sol) a.




```
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

(Intercept)
comp_pct    ***
---
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
```

From the graph, it is evident that the quadratic regression model (the pink line) fits the data better than the linear model in the range of higher years of education.
The quadratic regression model improves upon the linear model's underestimation of wages for individuals with higher education levels, reflecting the phenomenon of increasing marginal returns to education.
Although the R-squared values for both models are not high (both are below 25%), the quadratic model still performs slightly better and provides a better fit to the data.
Conclusion:
The quadratic model appears to fit the data better than the linear model.
If a more complex nonlinear relationship between education and wages needs to be captured, the quadratic model should be used.

b.

```
      fit       lwr      upr
81.92474 77.38223 86.46725
```

c.

```
> t_value
[1] 4.26536
> critical_value
[1] 2.499867
>
> if (t_value > critical_value) {
+   print("Reject H0: There is significant evidence that β2 > 0 at α = 0.01.")
+ } else {
+   print("Fail to reject H0.")
+ }
[1] "Reject H0: There is significant evidence that β2 > 0 at α = 0.01."
```

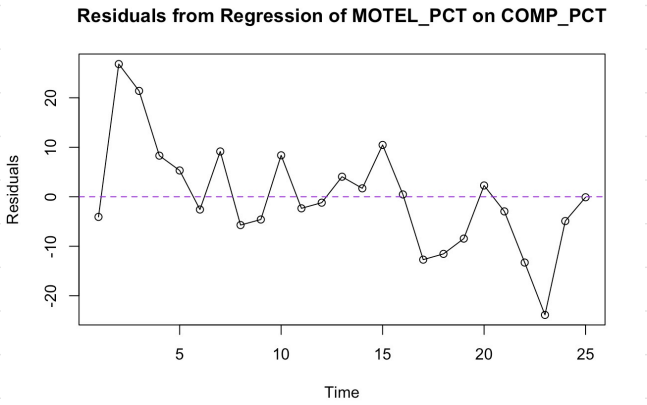
We reject H_0 and conclude that $\beta_2 > 0$ at the 1% significance level, indicating a significant positive relationship between COMP_PCT and MOTEL_PCT.

d.

```
> t_value2 <- (beta2 - 1) / se_beta2
> critical_value2 <- qt(1 - alpha / 2, df)
>
> t_value2
[1] -0.6677491
> critical_value2
[1] 2.807336
>
> if (abs(t_value2) > critical_value2) {
+   print("Reject H0: β2 is not equal to 1 at α = 0.01.")
+ } else {
+   print("Do not reject H0.")
+ }
[1] "Do not reject H0."
```

At the 0.01 significance level, the calculated t-value is -0.668, which lies within the acceptance region defined by the critical values ±2.807. Therefore, we do not reject the null hypothesis that β_2 equals 1. This suggests that the motel's occupancy rate increases proportionally with the competitor's occupancy rate, on a one-to-one basis.

e.




```
> subset_residuals <- residuals[motel$time >= 17 & motel$time <= 23]
> subset_residuals
       17         18         19         20         21
-12.707328 -11.543226  -8.456225   2.279673  -2.958191
       22         23
-13.293015 -23.875603
>
> sign(subset_residuals)
17 18 19 20 21 22 23
-1 -1 -1  1 -1 -1 -1
```

The residual plot shows some unusual features, particularly a noticeable downward trend after time period 17. During periods 17 to 23 (July 2004 to January 2005), most of the residuals are negative, indicating that the motel's actual occupancy rate was generally lower than what the model predicted. Out of these seven periods, six residuals are negative and only one is positive.