HW0310

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. Null Hypothesis (H_0) : $b_2 = 0$, no relationship between GDP and number of medals won.
Alternative Hypothesis (H_1) : $b_2 > 0$,
a positive relationship between GDP and number of medals won.

b. $t = \frac{b_2}{se(b_2)} = \frac{0.01309}{0.00215} = 6.088$ with the degrees of freedom $64-2=62$. Therefore, $t \sim t_{62}$.

c. Under H_0,
the expected value of b_2 is 0, so the test statistic follows a t-distribution centered at 0.
Under H_1,
b_2 is positive, so the expected value of the test statistic will be larger than 0 and thus be shifted to the right of the standard t-distribution.

d. Since $t_{0.01,62} = 2.388$, we reject H_0 if $t > 2.388$; otherwise, we fail to reject H_0.

e. Since $t = 6.088 > t_{0.01,62} = 2.388$, we reject the null hypothesis at the 1% significance level.
It means that there's strong evidence that GDP is positively related to the number of Olympic medals won.
The meaning of 1% significance level is the probability of rejecting H_0 when it is actually true is expected to be 1%.

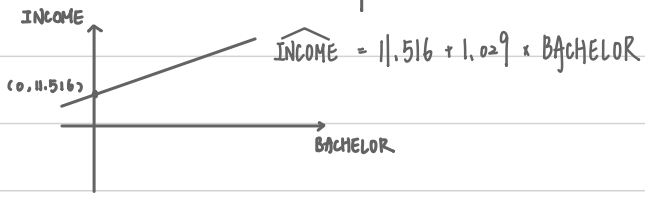
3.7 We have 2008 data on *INCOME* = income per capita (in thousands of dollars) and *BACHELOR* = percentage of the population with a bachelor's degree or more for the 50 U.S. States plus the District of Columbia, a total of $N = 51$ observations. The results from a simple linear regression of *INCOME*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

a. $\hat{a} = t_{intercept} \times se(a) = 4.31 \times 2.672 = 11.516$

b. The estimated relationship is : $\widehat{INCOME} = 11.516 + 1.029 \times BACHELOR$

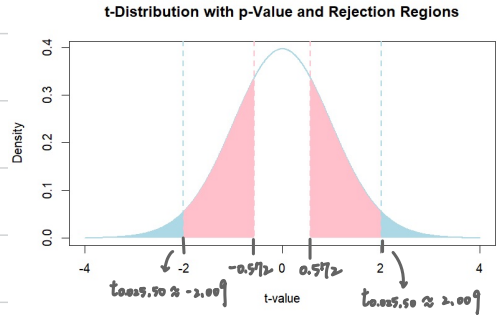


→ The slope is positive, so the relationship is increasing and positive.
→ Since it is a linear model, it increases at a constant rate.

c. $se(b_2) = \frac{b_2}{t_{slope}} = \frac{1.029}{10.75} = 0.0957$

d. $t = \frac{\hat{a} - 10}{se(a)} = \frac{11.516 - 10}{2.672} = 0.567$

e.



→ The blue region is the rejection region.
RR : $\{ t : t \le -2.009 \text{ or } t \ge 2.009 \}$.

f. $CI = b_2 \pm t_{critical} \times se(b_2) = (1.029 - 2.68 \times 0.0957,\ 1.029 + 2.68 \times 0.0957) = (0.772, 1.286)$
→ We are 99% confident the the true slope lies between 0.772 and 1.286

g. $H_0 : b_2 = 1$ → $t = \frac{b_2 - 1}{se(b_2)} = \frac{1.029 - 1}{0.0957} = 0.303 < t_{critical} = 2.010$
$H_1 : b_2 \neq 1$
⇒ fail to reject H_0.
⇒ It doesn't have strong evidence to conclude that the effect of education on income per capita is siginificantly different from 1.

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.

a. $\begin{cases} H_0: \beta_2 = 1.8 \\ H_1: \beta_2 > 1.8 \end{cases}$

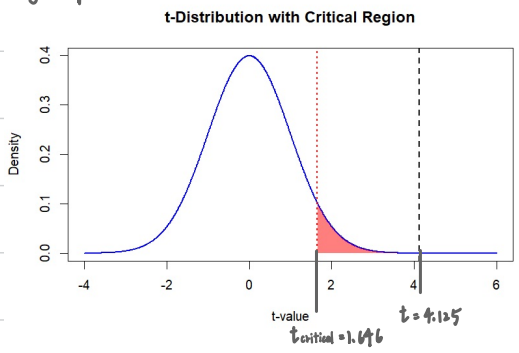
(1) test statistic : $t = \frac{2.46 - 1.8}{0.16} \approx 4.125$

(2) critical value : $t_{0.95,984} = 1.646$

(3) decision : Since $t = 4.125 > t_{0.95,984} = 1.646$, We reject H_0.

(4) conclusion: There's strong stastical evidence that education has a greater effort on wages in urban areas than in rural areas.

(5) graph :



b. (1) Given EDUC = 16, $\widehat{WAGE} = -4.88 + 1.8 \times 16 = 23.92$

(2) critical value : $t_{0.975,212} \approx 1.97$

(3) $se(\widehat{WAGE}) = \sqrt{se(\beta_1)^2 + 16^2 se(\beta_2)^2 + 2 \times 16 \times cov(\beta_1, \beta_2)}$

$= (3.29^2 + 16^2 \times 0.24^2 + 2 \times 16 \times (-0.761))^{\frac{1}{2}} \approx 0.833$

(4) confidence interval : $CI = 23.92 \pm 0.833 \times 1.97 = (22.279, 25.561)$

c. (1) Given EDUC = 16, $\widehat{WAGE} = -10.76 + 2.46 \times 16 = 28.6$

(2) critical value : $t_{0.975,\,2132} \approx 1.97$

(3) $se(\widehat{WAGE}) = \sqrt{se(\beta_1)^2 + 16^2 se(\beta_2)^2 + 2 \times 16 \times cov(\beta_1, \beta_2)}$

$= (2.27^2 + 16^2 \times 0.16^2 + 2 \times 16 \times (-0.345))^{\frac{1}{2}} \approx 0.816$

(4) confidence interval : $CI = 28.6 \pm 0.816 \times 1.97 = (26.990,\ 30.208)$

(5) Comparism : Since $1.1035 > 0.816$, urban is narrower.

→ It is plausible, since the urban sample size is larger, leading to a more precise estimate with a smaller standard error.

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

→ test statistic : $t = \frac{-4.88 - 4}{3.29} = -2.699$

→ critical value : $t_{0.01,\,2132} = -2.326$

→ Since $t = -2.699 < t_{0.01,\,2132} = -2.326$, we reject H_0.

→ There's sufficient evidence to conclude that $b_1 < 4$.

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.

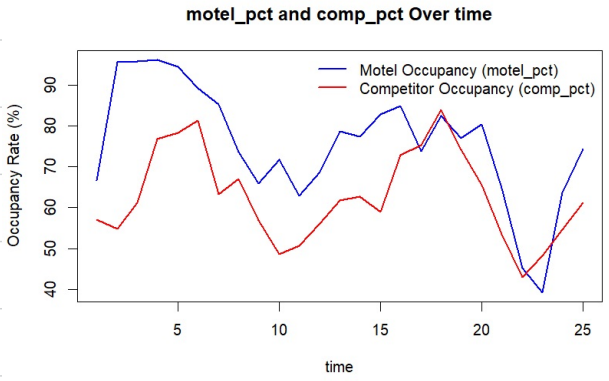
b. Construct a 90% interval estimate of the expected occupancy rate of the motel in question, *MOTEL_PCT*, given that $COMP\_PCT = 70$.

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)?

a.



→ They tend to move together in a similar pattern, suggesting a positive correlation between the occupancy rates of the motel and its competitors.

→ MOTEL_PCT seems to have the higher occupancy rates.

→

```
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
```

⇒ $MOTEL\_PCT = 21.4 + 0.8646\, COMP\_PCT + e$

→ confidence level : CI = (0.4453, 1.284)

→ Yes, we have estimated the association relatively precisely.
To be more specific, the p-value = 0.000296 is extremely low.

b. confidence level : CI = (77.3823 , 86.46725)

c. $\begin{cases} H_0 : \beta_2 \le 0 \\ H_1 : \beta_2 > 0 \end{cases}$

→ test statistic : $t = \frac{\beta_2}{se(\beta_2)} = 4.26536$

→ critical value : $t_{0.99,23} = 2.49987$

→ rejection region : $\{ t : t > 2.49987 \}$

→ Since $t = 4.26536 > t_{0.99,23} = 2.49987$, it falls within the rejection region. We reject H_0, meaning there's a statistically significant positive effect that COMP_PCT has on MOTEL_PCT at the 1% level.

d. $\begin{cases} H_0 : \beta_2 = 1 \\ H_1 : \beta_2 \neq 1 \end{cases}$

→ test statistic : $t = \frac{\beta_2 - 1}{se(\beta_2)} = -0.66775$

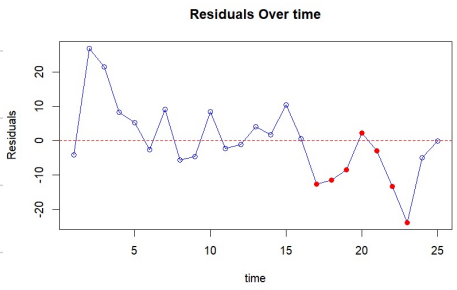
→ critical value : $t_{0.99,23} = 2.80734$

→ rejection region : $\{ t : |t| > 2.80734 \}$

→ Since $|t| = 0.66775 < t_{0.99,23} = 2.80734$, it does NOT fall within the rejection region. We fail to reject H_0, meaning β_2 is significantly different from 1.

→ If H_0 is true, then for evey 1% increase in COMP_PCT, MOTEL_PCT would increase by exactly 1%.

e.



→ In the period of 17-23, most of these residuals appear negative. The model overestimated MOTEL_PCT during these months.