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

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

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.

a. Null hypothesis: $H_0: \beta_2 = 0$, Alternative hypothesis: $H_1: \beta_2 > 0$

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

b. $t = \dfrac{b_2 - \beta_2}{se(b_2)} \sim t_{(N-2)} \left(t = \dfrac{0.01309 - 0}{0.00215} \approx 6.088 \right) \Rightarrow t = \dfrac{b_2}{se(b_2)} \sim t_{(62)}$ #

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?]

c. The center of the t-distribution shifs to the right if the alternative hypothesis 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?

d. Reject the null hypothesis and acept the alternative:

$t \geq 2.388$ (Rcode: qt(0.99, 62))

Fail to reject the null hypothesis:

$t < 2.388$

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?

e. $t = \dfrac{0.01309 - 0}{0.00215} \approx 6.0884 > 2.388$ We reject the null hypothesis and accept the alternative hypothesis, and it shows that there is positive relationship between the number of medals won and GDP.

$\alpha = 1\%$ means Type I error (the probability that we reject H_0 when it is true)

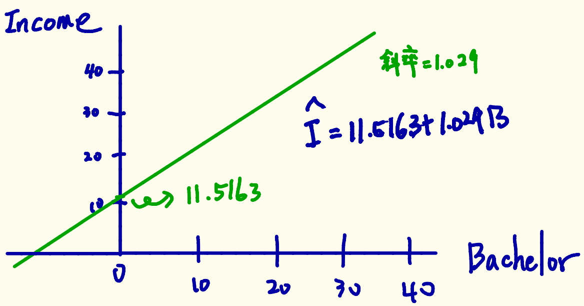
3.7 We have 2008 data on *INCOME* = income per capita (in thousands of dollars) and *BACHELOR* = percentage of the population with a bachelor's degree or more for the 50 U.S. States plus the District of Columbia, a total of $N = 51$ observations. The results from a simple linear regression of *INCOME*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

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

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

$$\frac{a}{2.672} = 4.31 \Rightarrow a = 4.31 \times 2.672 \approx 11.5163 \ \#$$

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?



It is a <u>positive relationship</u> (m = 1.029) and <u>increasing</u> at a <u>constant rate</u>. #

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

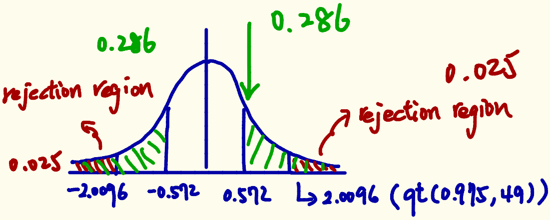
$$\frac{1.029}{c} = 10.75 \Rightarrow c = \frac{1.029}{10.75} \approx 0.0957 \ \#$$

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

$$H_0 : \beta_1 = 10$$

$$t = \frac{11.5163 - 10}{2.672} \approx 0.5674775 \approx 0.5675 \sim t(49)$$

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 $\alpha = 0.05$.



rejection region = $\{t \mid t \le -2.0096$ or $t \ge 2.0096\}$

Thus, we can't reject null hypothesis.

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

$$1.029 \pm \overset{qt(0.995,49)}{t_{(0.995,49)}} \times 0.0957$$
$$= 1.029 \pm 2.68 \times 0.0957$$
$$= [\,0.7725\ ,\ 1.2855\,] \ \#$$

在99%的信賴水準下, Bachelor 對 Income 的影響(β_2)會落於此區間。

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.

$$\begin{cases} H_0: \beta_2 = 1 \\ H_1: \beta_2 \ne 1 \end{cases} \qquad t = \frac{1.029 - 1}{0.0957} \approx 0.303_{\#} < 2.0096\,(t_{0.975,49})$$

we can't reject $H_0: \beta_2 = 1$, This suggests that the effect of the percentage of the population with a bachelor's degree on income per capita is statistically not 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.

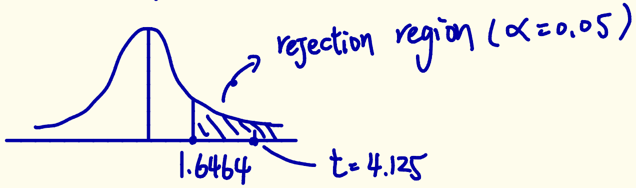
$$\text{Urban} \qquad \begin{aligned} \widehat{WAGE} &= -10.76 + 2.46EDUC, \quad N = 986 \\ (\text{se}) \quad & \quad\;\; (2.27) \quad (0.16) \end{aligned}$$

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

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.

a. $\begin{cases} H_0 : \beta_2 = 1.8 \\ H_1 : \beta_2 > 1.8 \end{cases}$ $\quad t = \dfrac{2.46-1.8}{0.16} = 4.125 > t_{0.95,984}$ (qt(0.95,984)) ≈ 1.6464 #

Thus, we can reject H_0 ($\beta_2 = 1.8$) and accept the H_1 ($\beta_2 > 1.8$).
It means when education increase one year, hourly wage rate increase more 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.

$$\widehat{WAGE} = -4.88 + 1.80 \times 16 = 23.92$$

$$t_{(0.025,212)} = qt(0.025,212) = -1.9712$$

$$\begin{aligned} \text{C.I.} &= \mu \pm t_{(0.025,212)} \times se \\ &= 23.92 \pm (-1.9712) \times 0.833 \\ &\Rightarrow [22.278 \;,\; 25.562] \ \# \end{aligned}$$

$$\left(\begin{aligned} se(\widehat{WAGE}) &= \sqrt{se(B_0)^2 + (EDUC)^2 se(\beta_1)^2 + 2\,EDUC\,cov(\beta_0, \beta_1)} \\ &= \sqrt{3.29^2 + 16^2 \times 0.24^2 + 2 \times 16 \times (-0.761)} = 1.1035 \end{aligned} \right)$$

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.

$educ = 16 \quad \widehat{WAGE} = -10.76 + 2.46 \times 16 = 28.6$

$se(WAGE) = \sqrt{2.27^2 + 16^2 (0.16)^2 + 2 \cdot 16 \cdot (-0.345)} = 0.8164$

$t_{(0.025, 984)} = qt(0.025, 984) = -1.96$

C.I. $\begin{cases} 28.6 + 1.96 \times 0.8164 = 30.20 \\ 28.6 - 1.96 \times 0.8164 = 27.00 \end{cases}$

$95\%\ CI = [27.00,\ 30.20]$

⇒ The interval estimate for the urban regression narrower than that for rural regression because there are more samples in urban regression, which cause $t_{(0.975, N)}$ is smaller than rural regression and the standard error for urban regression is smaller than rural regression.

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.

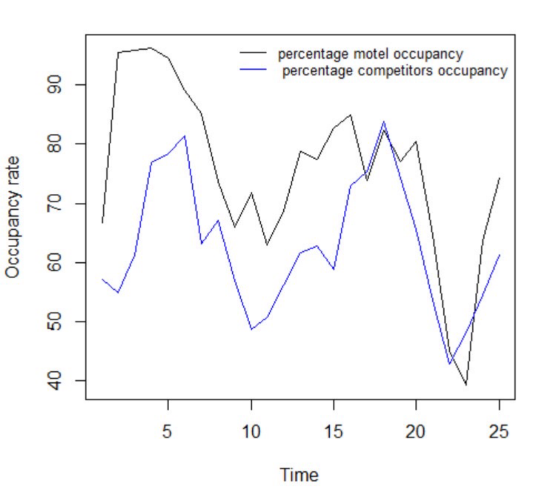
$\begin{cases} H_0: \beta_1 \geq 4 \\ H_1: \beta_1 < 4 \end{cases} \quad t = \dfrac{-4.88 - 4}{3.29} \approx -2.6991 < \underbrace{t_{0.01, 212}}_{qt(0.01, 212)} = -2.3441$

we can reject $H_0: \beta_1 \geq 4$, Accept $\beta_1 < 4$, at 1% level

(it falls in the rejection region)

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.



⇒ The occupancy rates go up and down over time.

MOTEL_PCT is higher than COMP_PCT most of the time.

They move together.

$$\widehat{MOTEL\_PCT} = 21.4 + 0.8646\,COMP\_PCT$$
$$(se) \quad (12.9069) \quad (0.2027)$$

```
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 ***
```

```
              2.5 %    97.5 %
comp_pct 0.4452978 1.283981
```

→ 95% C.I. of β_2 is $[0.445, 1.284]$

we don't estimate the association relatively precisely since C.I. is width.

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

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

90% interval estimate $= [77.382, 86.467]$ #

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

$$t = \frac{0.8646}{0.2027} = 4.265_{\#} > t_{(0.99,\,23)} = 2.499867$$

We can reject $H_0: \beta_2 \le 0$, Accept $H_1: \beta_2 > 0$

Thus, there is a positive relationship between MOTEL-PCT and COMP_PCT.

Rejection region: $\{t \mid t \geq 2.5\}_{\#}$

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.

$$t值 = \frac{b_2 - 1}{se(b_2)} = \frac{0.8646 - 1}{0.2027} = -0.6679822 \approx -0.67_{\#} > t_{(0.005,\,23)} = -2.807336$$

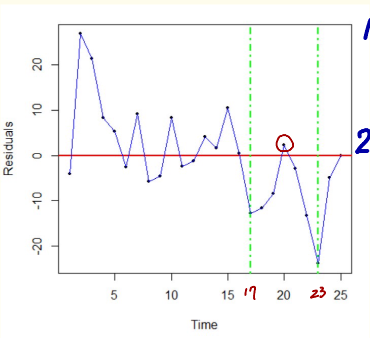
qt(0.005,23)

Rejection region: $\{t \mid t \geq 2.8073 \text{ or } t \leq -2.807\}_{\#}$

we can't reject the null hypothesis ($H_0: \beta_2 = 1$) 沒有足夠證據表明 β_2與1顯著不同

→ The sample data are consistent with the conjecture that when COMP_PCT increase 1%, MOTEL_PCT will increase 1%, holding all else constant.

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. 在0~15，大多數殘差都是正的，暗指 MOTEL-PCT 是高於模型預期值

2. 而17-23，除了一点，其他殘差都是負的，指出 MOTEL-PCT是低於模型預期值