Homework 6  
Statistical Inference

**PREPARED BY**

Mohsen Fayyaz - 810100524

Winter 2022

**Table of Contents**

[**Q1**](#_tbfpyld5fvy5) **3**

[Answer to the following questions:](#_s3bv1as6056v) 3

[**2**](#_w5lvuohxq6rz) **6**

[In the table below, the 𝑥𝑖 column shows scores on the aptitude test. Similarly, the 𝑦𝑖 column shows statistics grades. Find the Regression Equation.](#_f7nmrkbcmnnc) 6

[**3**](#_ttkoak8p4qny) **7**

[Answer the following questions:](#_5d54obp50rnt) 7

[**4**](#_3oct7fj70zih) **9**

[(R) The openintro package contains a dataset called absenteeism that consists of data on 146 schoolchildren in a rural area of Australia. Spend some time reading the help file of this dataset. We are interested in seeing if the ethnicity (aboriginal or not), sex (male or female), and learning ability (average or slow) of the children affects the number of days they are absent from school.](#_5p9o9n9eyy29) 9

[**5**](#_7z1jk2k8frhz) **12**

[Determine if the following statements are true or false. If false, explain why.](#_4d3owx6ir4fz) 12

[**6**](#_sd2ubpcdvdtl) **13**

[The following regression output is for predicting annual murders per million from percentage living in poverty in a random sample of 20 metropolitan areas.](#_h2uib3wl0d6j) 13

[**7**](#_wi4clkphf8wx) **15**

[A traveler thinks that the time it takes them to commute is different based on which day of the week it is. For a few weeks, they record their commute time each day, Monday through Friday. Test their claim at the 5% level.](#_uyenfnwb5w38) 15

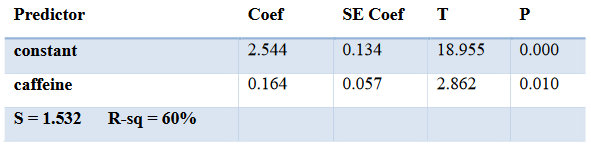
[**8**](#_d9hzawwsqtos) **16**

[(R) The built-in “state.77” dataset in R includes information about 50 states in US. Take life expectancy as the response and the remaining variables as predictors. Using this dataset, answer the following questions:](#_l3clm5jwsz8k) 16

# Q1

## Answer to the following questions:

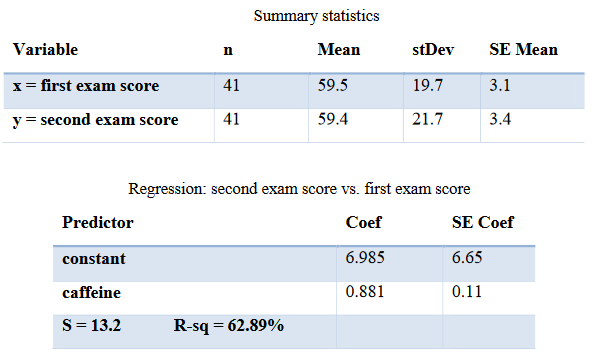
a. Ali is studying the correlation between hours spent studying and caffeine consumption among students at his school. During a given week, he randomly selects 20 students at his school and records their caffeine intake (mg) and studying time. Here is computer output from a least-squares regression analysis on his sample:



Assume that all conditions for inference have been met. Which of these is a 95% confidence interval for the slope of the least-squares regression line?

SE coef for caffeine which is 0.057 is the confidence interval for the slope of the least-squares regression line.   
df = n - 2 = 18  
t\*(95%, 18)=qt(0.025, df=18, lower.tail = FALSE)=2.101  
Thus, the final interval is

b. Elvin compared the scores of a random sample of 41 students on a first exam in a certain class with their subsequent scores on the second exam. Here is computer output on the sample data:

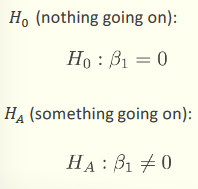


i. What conditions should be met for the above inferences?

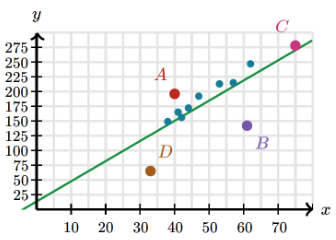
1) Linearity:   
Relationship between the explanatory (x) and the response (y) variable should be linear.

2) Nearly normal residuals:  
residuals should be nearly normally distributed

3) Constant variability:  
variability of points around the least squares line should be roughly constant

ii. Write an appropriate test statistic for testing the null hypothesis that the population slope in this setting is 0?  


c. Sara was curious which points in the scatterplot below were most influential in terms of the coefficient of determination, 𝑟2.



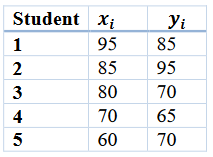
Which point, if removed, would cause 𝑟2 to decrease the most? Why?

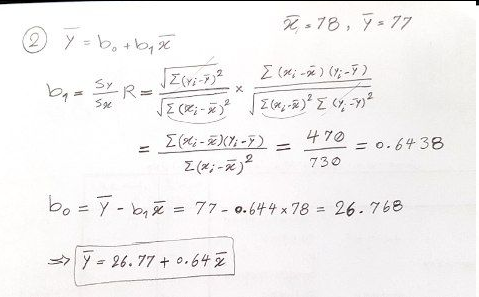
Removing point C will decrease r2 the most.

This is because C is the closest to the regression line. Therefore, removing it will reduce the association and correlation, thus decreasing the r2 the most.

# 2

## In the table below, the 𝑥𝑖 column shows scores on the aptitude test. Similarly, the 𝑦𝑖 column shows statistics grades. Find the Regression Equation.





# 3

## Answer the following questions:

a. Each morning, Arianna runs. For a random sample of runs, she tracked the temperature (in degrees Celsius) and the distance run (in kilometers). The temperatures were negatively correlated with the distances. A 95% confidence interval for the slope of the regression line was (-0.02, 0.12). Arianna wants to use this interval to test 𝐻0: 𝛽 = 0 vs 𝐻𝑎: 𝛽 ≠ 0 at the 𝛼 = 0.05 level of significance. Assume that all conditions for inference have been met. Which of these is the most appropriate conclusion about the relationship between temperature and distance for Arianna’s runs?

i. Reject 𝐻0. Ariana can’t conclude a linear relationship between temperature and distance.

ii. Fail to reject 𝐻0. Ariana can’t conclude a linear relationship between temperature and distance.

iii. Reject 𝐻0. This suggests a linear relationship between temperature and distance.

iv. Fail to reject 𝐻0. This suggests a linear relationship between temperature and distance.

(ii) is correct. As the 95% confidence interval (-0.02, 0.12) contains the null hypothesis 𝐻0: 𝛽 = 0, we cannot reject H0. And this means that we cannot infer a significant linear relationship between temperature and distance. H0 was the claim that the explanatory variable is not a significant predictor of the response variable, i.e. no relationship → slope of the relationship is 0.

b. Biologists observed a curved relationship between the average heart rate and life expectancy of several mammal species in a large sample. The biologist took the natural logarithm for both variables and observed a linear relationship in the transformed data. The least-squares regression equation for the transformed data is shown here, where life expectancy (LE) is in years and heart rate (HR) is in beats per minute.

𝑙𝑛(𝐿𝐸)̂ = 6.33 − 0.78 𝑙𝑛(𝐻𝑅)

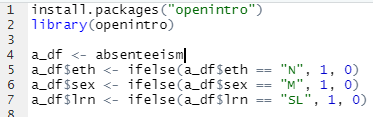
What is the predicted life expectancy of a mammal species whose average heart rate is 60 beats per minute according to this model? (Round your answer to the nearest year.)

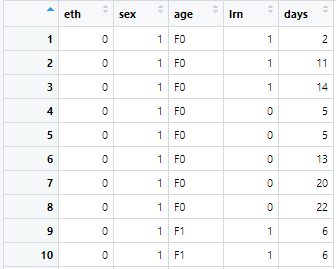
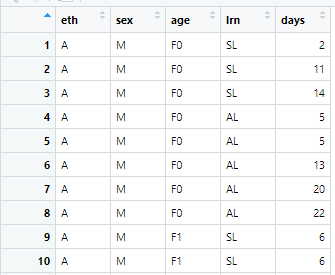
= 23 years

# 4

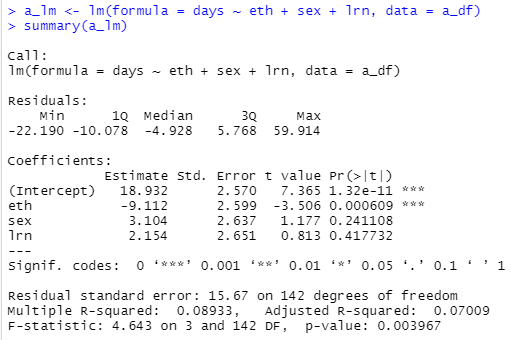
## (R) The openintro package contains a dataset called absenteeism that consists of data on 146 schoolchildren in a rural area of Australia. Spend some time reading the help file of this dataset. We are interested in seeing if the ethnicity (aboriginal or not), sex (male or female), and learning ability (average or slow) of the children affects the number of days they are absent from school.

a. Convert the Eth, Sex, and Lrn variables to binary variables. One way to do this is with the function ifelse(). You should construct them so that:  
i. Eth = 1 if the student is not aboriginal and Eth = 0 if the student is aboriginal;  
ii. Sex = 1 if the student is male and Sex = 0 if the student is female;  
iii. Lrn = 1 if the student is a slow learner and Lrn = 0 is the student is an average learner.





b. Fit a linear model to the data with Days as the dependent variable and the three variables mentioned in (1) as explanatory variables.



c. Write the fitted model out using mathematical notation. Make sure you define the variables you use. Interpret all of the coefficients (including the intercept) in context.

Days is the number of days they are absent from school  
Eth is ethnicity (aboriginal[0] or not[1])  
Sex is male[1] or female[0]  
Lrn is learning ability (average[0] or slow[1])

* Intercept means that a student who is aboriginal, female, and slow learner is expected to be absent for 18.932
* Eth coefficient means that all else held constant, if the ethnicity is not aboriginal the model predicts the absent days to be less on average by 9.112 days.
* Sex coefficient means that all else held constant, if the student is male the model predicts the absent days to be more on average by 3.104 days.
* Lrn coefficient means that all else held constant, if the student is a slow learner the model predicts the absent days to be more on average by 2.154 days.

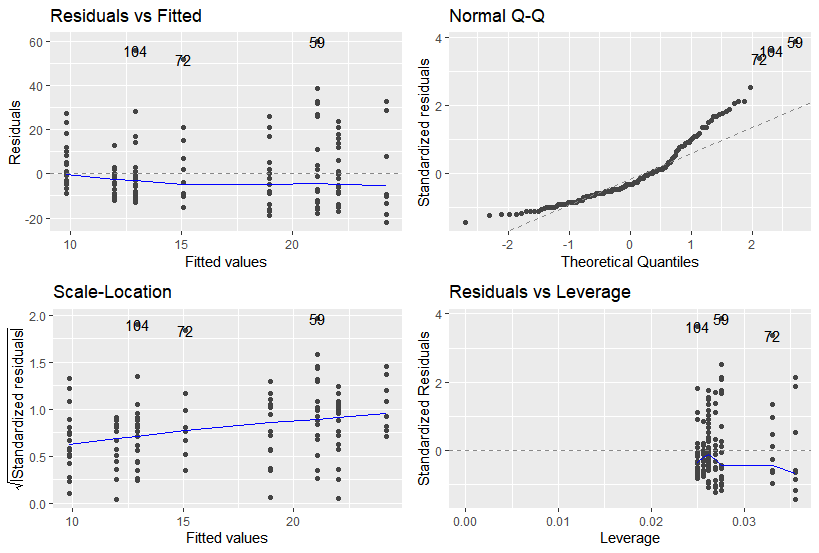
d. Find and interpret the adjusted 𝑅2 value for this model.



Based on our analysis adjusted R2 is 0.07 which means that our predictors (eth,sex,lrn) can explain the 7% of the variability of absent days.

e. Create a residual plot. Describe what you see in the residual plot. Does the model look like a good fit?





We see in Residuals w/ fitted values that at the tails the residuals distribution gets more different than normal distribution and this can also be observed in the qq plot. Therefore, our model is a good fit for the middle parts, but it is not as good of a predictor at tails.

# 5

## Determine if the following statements are true or false. If false, explain why.

a. By adding an explanatory variable to an existing MLR model if the variable is not a meaningful predictor, R2 will decrease and adjusted R2 will stay about the same.

False:  
When any variable is added to the model 𝑅2 increases. However, if the added variable doesn’t really provide any new information, or is completely unrelated, adjusted 𝑅2 does not increase.

b. A correlation coefficient of -0.90 indicates a stronger linear relationship than a correlation of 0.5.

True:  
The magnitude of the correlation coefficient determines the strength of a linear relationship. The positive or negative sign only indicates the positive or negative relationship not the strength. Thus, -0.90 indicates a stronger linear relationship than a correlation of 0.5.

c. Correlation is a measure of any association between any two variables.

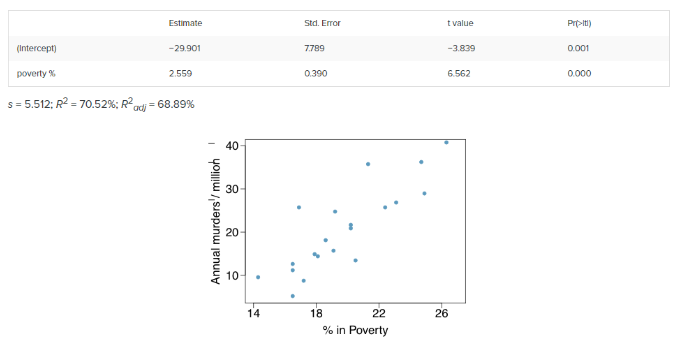
False:  
Correlation describes the strength of the linear association between two variables.

d. If predictors are collinear, then removing one variable will have no influence on the point estimate of another variable’s coefficient.

False:  
Two predictor variables are said to be collinear when they are correlated with each other. By removing one variable, we are removing a column of data which would change the calculation of another variable’s point estimate.

# 6

## The following regression output is for predicting annual murders per million from percentage living in poverty in a random sample of 20 metropolitan areas.



a. Write out the linear model.

Murder = y, poverty = x

b. Interpret the intercept, slope, R2.

Intercept:   
murder rate is expected to be -29.901 in an area with no poverty (x=0). This is clearly not true and inference on the intercept is rarely done.

Slope:  
For a unit increase (1%) in poverty we expect to see 2.559 increase in murders per million

R2:  
Percentage living in poverty explains 70.52% of the variability in murder rate in this city.

c. Calculate the correlation coefficient.

The magnitude is from R2. However, for the sign of the correlation, we look at the plot and as the association seems to be positive we put + before correlation magnitude.

d. What are the hypotheses for evaluating whether poverty percentage is a significant predictor of murder rate?

𝐻0 (nothing going on): The explanatory variable (poverty percentage) is not a significant predictor of the response variable (murder rate), i.e. no relationship → slope of the relationship is 0.

𝐻𝐴 (something going on): The explanatory variable (poverty percentage) is a significant predictor of the response variable (murder rate), i.e. relationship → slope of the relationship is different than 0.

e. State the conclusion of the hypothesis test from part d in context of the data.

Based on the table, the p-value is approximately 0. Hence, we reject the null hypothesis H0. Therefore, the data provides convincing evidence that poverty percentage is a significant predictor of murder rate.

f. Calculate a 95% confidence interval for the slope of poverty percentage, and interpret it in context of the data.

We are 95% confident that a unit increase (1%) in poverty percentage is epeected to increase murder rate by 1.74 to 3.38.

g. Do your results from the hypothesis test and the confidence interval agree? Explain.

Yes. The confidence interval 1.74 to 3.38 does not include 0, thus both reject H0.

# 7

## A traveler thinks that the time it takes them to commute is different based on which day of the week it is. For a few weeks, they record their commute time each day, Monday through Friday. Test their claim at the 5% level.

a. What kind of test should we conduct?

ANOVA  
Partitioning the variability in 𝑦 to explained and unexplained variability requires analysis of variance (ANOVA).

b. What are the hypotheses?

c. Complete the missing spaces in partially filled table.

| Source | DF | Sum of  Squares | Mean Square | F Value | Prob. |
| --- | --- | --- | --- | --- | --- |
| Day(Groups) | 5-1=**4** | 14.28 | **3.57** | **3.37** | pf(3.37, df1 = 4, df2 = 15, lower.tail = FALSE) = **0.0371221** |
| Error(Residuals) | 19-4=**15** | 30.2-14.28=**15.92** | **1.06** |  |  |
| Total | 19 | 30.2 |  |  |  |

d. What is the correct decision?

As the p-value is 0.037<0.05 we reject the null hypothesis H0.

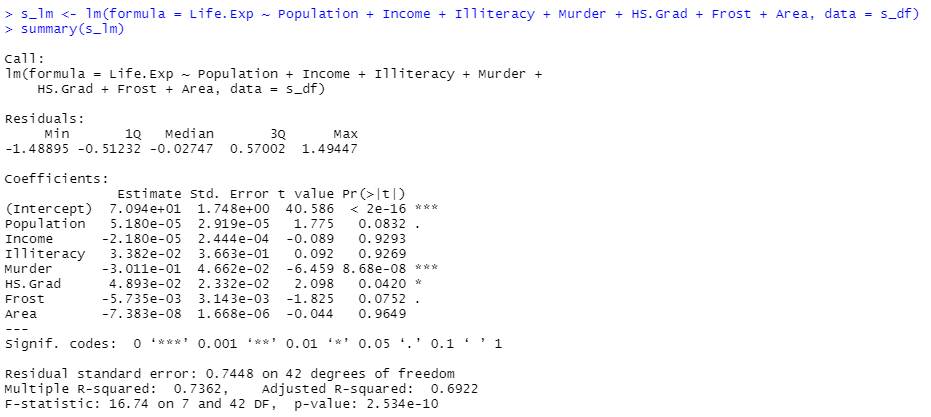
e. What is the appropriate conclusion/interpretation?   
The data provides convincing evidence that commute time and day of the week are associated.

# 8

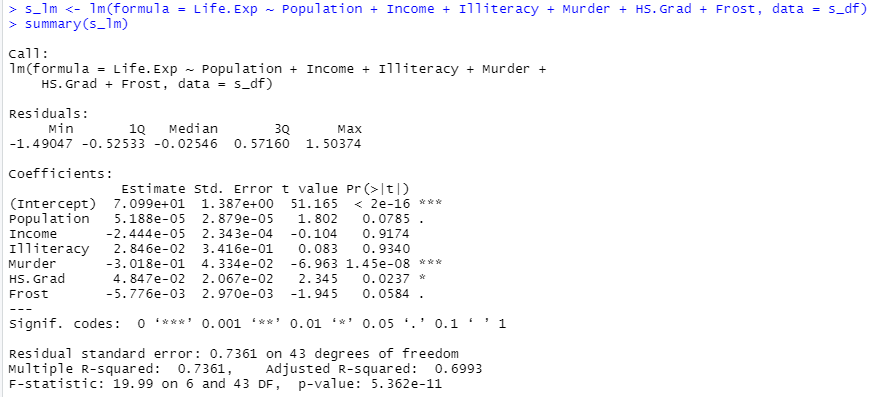
## (R) The built-in “state.77” dataset in R includes information about 50 states in US. Take life expectancy as the response and the remaining variables as predictors. Using this dataset, answer the following questions:

a. Using backward selection and p-value as the criterion, find a multiple (or simple) linear regression model which is the best for predicting life expectancy using others. Show each step in your report.

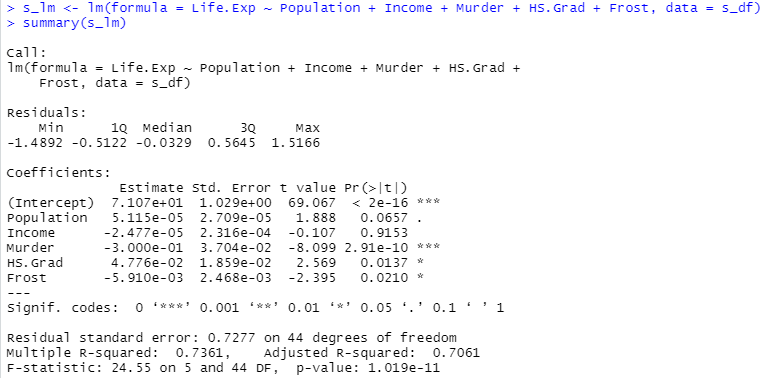
Backwards Elimination - p-value



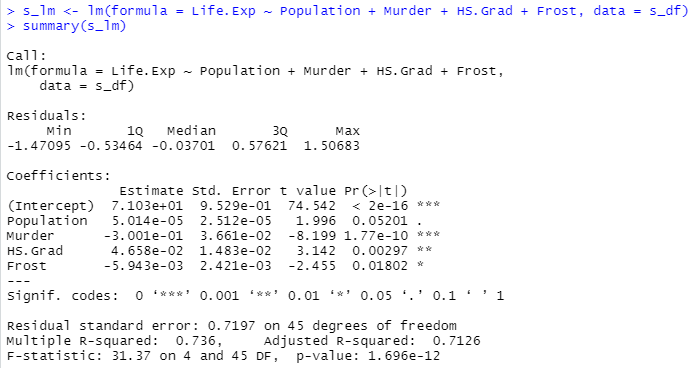
Highest p-value is for Area therefore we drop it.



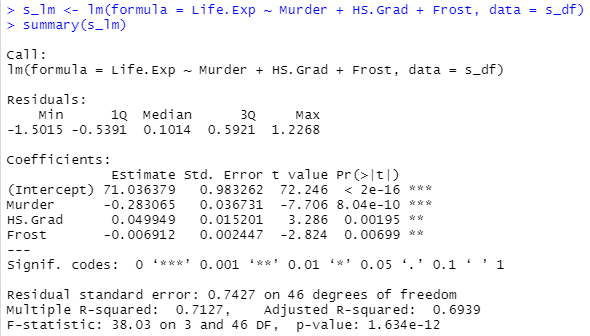
Highest p-value is for Illiteracy therefore we drop it.



Highest p-value is for Income therefore we drop it.

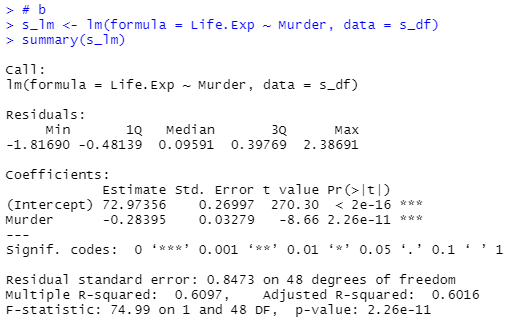


Highest p-value is for Population therefore we drop it.



Now all three predictors left have a p-value larger than 0.05, thus all variables left in the model are significant.

b. Find the simple linear model of data. (Assume X = Murder and Y = Life expectancy).

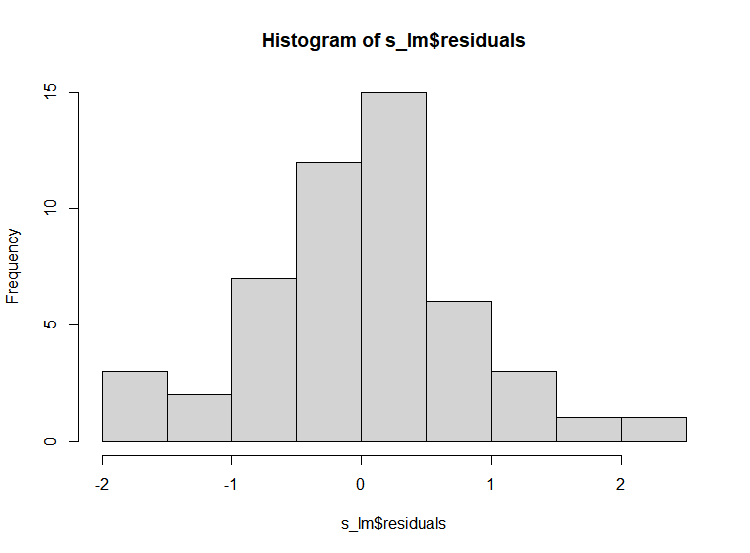


Y = 72.97 - 0.28\*X

This means that for a unit increase in murder, we expect the life expectency to decrease by 0.28 years. Also, the intercept can be interpreted as if there is no murder, we expect the life expectancy to be 72.97 years.

c. Plot histogram of residuals and find mean and sd for part (b).





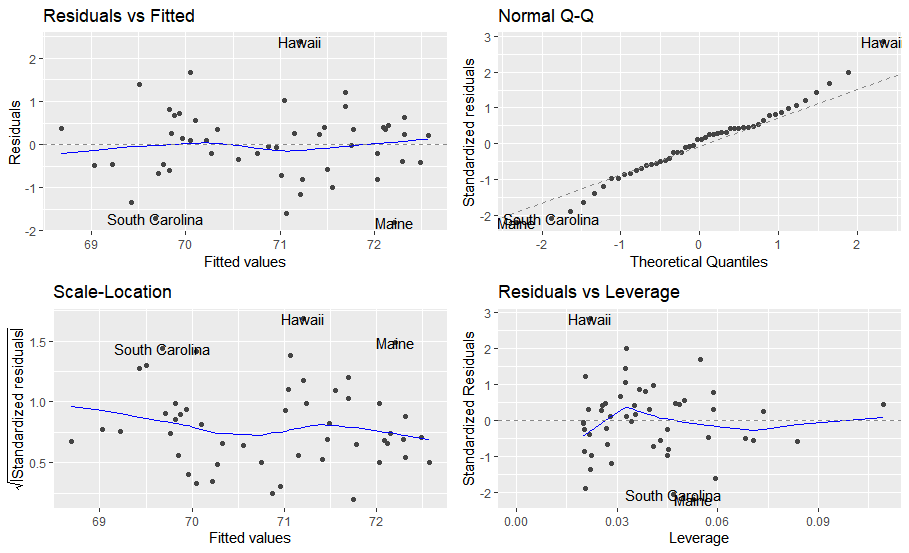


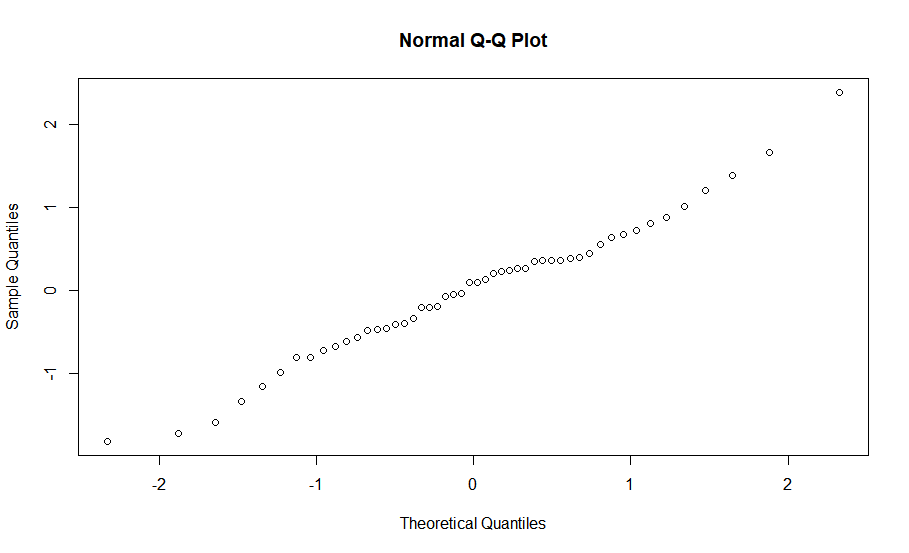
Y = 72.97 - 0.28\*X

This means that for a unit increase in murder, we expect the life expectency to decrease by 0.28 years. Also, the intercept can be interpreted as if there is no murder, we expect the life expectancy to be 72.97 years.

d. Plot QQ-plot of residuals versus a zero-mean normal distribution with sd of part (b). Do you observe that residuals are nearly normal?







As we see in the plots, our predictor is a good fit for the middle parts where the distribution of residuals is close to normal distribution; however, we do not have a good predictor at the tails, as the distribution of the residuals gets further from normal distribution.