STAT 308 – Take Home Exam 2

This take home exam is worth a total of 50 points, and counts for half of your Exam 1 grade, along with the in class portion of the exam. For the problems in which calculations are needed, please include your `R` code with your answers. Turn in this exam to Sakai by Thursday, November 17 at 2:30 pm.

1. [3 pts each] For each of the following problems with a sample size of n, calculate an appropriate test statistic and p-value for a linear model with the given categorical variable with p possible categories as the lone explanatory variable in the model. Assume there are no other explanatory variables in the model.

   a. $n = 50$, $p = 4$, $SSM = 150.32$, $SSE = 230.44$.

```
SSM <- 150.32
SSE <- 230.44
n <- 50
p <- 4
F <- (SSM/(p-1))/(SSE/(n-p))
p.val <- pf(F,p-1,n-p,lower.tail=FALSE)
```

   $F = 10.00$, $p = 3.432e-5$

   b. $n = 15$, $p = 3$, $SSM = 77.3$, $SSE = 125.44$.

```
SSM <- 77.3
SSE <- 125.44
n <- 15
p <- 3
F <- (SSM/(p-1))/(SSE/(n-p))
p.val <- pf(F,p-1,n-p,lower.tail=FALSE)
```

   $F = 3.697$, $p = 0.0561$

   c. $n = 60$, $p = 6$, $SSM = 50.32$, $SSE = 250.78$.

```
SSM <- 50.32
SSE <- 250.78
n <- 60
p <- 6
F <- (SSM/(p-1))/(SSE/(n-p))
p.val <- pf(F,p-1,n-p,lower.tail=FALSE)
```

   $F = 2.167$, $p = 0.0713$

   d. $n = 75$, $p = 5$, $SSM = 3.22$, $SSE = 9.87$.

```
SSM <- 3.22
SSE <- 9.87
n <- 75
p <- 5
F <- (SSM/(p-1))/(SSE/(n-p))
p.val <- pf(F,p-1,n-p,lower.tail=FALSE)
```

$F = 5.709$, $p = 0.000491$

e. $n = 28$, $p = 2$, $SSM = 60.44$, $SSE = 514.32$.

```
SSM <- 60.44
SSE <- 514.32
n <- 28
p <- 2
F <- (SSM/(p-1))/(SSE/(n-p))
p.val <- pf(F,p-1,n-p,lower.tail=FALSE)
```

$F = 3.055$, $p = 0.09227$

2. For this problem, consider the `mtcars` dataset available in base `R`. Use the command "?mtcars" to obtain information about the overall dataset, as well as the variables contained within it.

Suppose we would like to be able to predict a car's fuel efficiency in miles per gallon based on the car's horsepower.

a. [3 pts] Perform a formal hypothesis test to determine if horsepower is linear related to the car's fuel efficiency. Be sure to include all pieces of information needed to perform a hypothesis test. Assume $\alpha = 0.05$.

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

```
mod <- lm(mpg ~ hp,mtcars)
summary(mod)
```

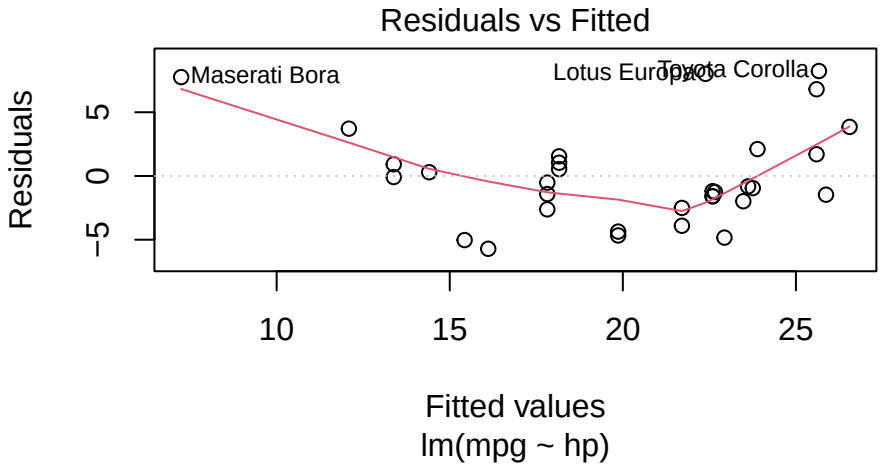
```
##
## Call:
## lm(formula = mpg ~ hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7121 -2.1122 -0.8854  1.5819  8.2360
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
## hp          -0.06823    0.01012  -6.742 1.79e-07 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.863 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

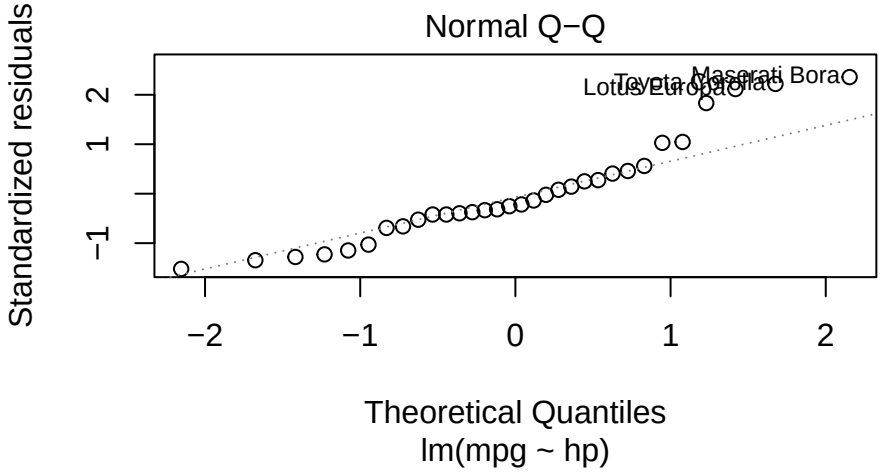
$F = 45.46$, $p = 1.788e-7$. We reject H_0 and conclude that there is statistically significant evidence of a linear relationship between horsepower and fuel efficiency in miles per gallon.

b. [2 pts] Check if the assumptions of homoscedasticity and normally distributed are violated. If they are not met, suggest how we can adjust our model so that these assumptions are not violated.

```
plot(mod,1)
```




```
plot(mod,2)
```



There is signficant curvature in the residual plot and the points in the upper tail of the QQ plot fall well above the 45-degree line, suggesting that both assumptions of homoscedasticity and normally distributed residuals are violated. Based on the curvature in the residual plot, perhaps adding a quadratic term to the model would be beneficial.

c. Suppose that, after discussion with fellow researchers, we decide to add a quadratic term for horsepower to our model. Perform a formal hypothesis test by answering the following questions.

- i. [2 pts] Write the appropriate null and alternative hypotheses.

Reduced Model: $\hat{mpg} = \beta_0 + \beta_1 hp$ Full Model: $\hat{mpg} = \beta_0 + \beta_1 hp + \beta_2 hp^2$

$$H_0 : \beta_2 = 0$$
$$H_a : \beta_2 \neq 0$$

- ii. [1 pt] What are the sums of squares that are added to the model when we add the quadratic term?

```
mod_full <- lm(mpg ~ hp + I(hp^2),mtcars)
anova(mod,mod_full)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ hp
## Model 2: mpg ~ hp + I(hp^2)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 447.67
## 2     29 274.63  1    173.04 18.273 0.0001889 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sums of squares added to the model is 173.04.

- iii. [1 pt] What are the error sums of squares for the polynomial model that includes both a linear and quadratic term for horsepower?

Sums of squares for the model with both a linear and quadratic term for horsepower is 274.63.

- iv. [2 pts] What is the F statistic and p-value?

$F = 18.273$, $p = 0.0001889$

- v. [2 pts] State your decision and conclusion in the context of the problem. We reject H_0 and conclude that adding a quadratic term for horsepower to a model that already includes a linear term for horsepower significantly improves the predictive ability of fuel efficiency (in mpg).

3. Consider the `lifeexp` dataset on the course webpage which contains information on countries from the year 2015. The variables contained in the dataset are defined as follows:

- `Status`: Development status of the country
- `Life.Expectancy`: Country's average life expectancy in years
- `Adult.Mortality`: Number of deaths of people between 15 and 60 per 1000 population
- `infant.deaths`: Number of infant deaths per 1000 population
- `Measles`: Number of measles cases per 1000 population
- `BMI`: average BMI for entire population
- `under.five.deaths`: Number of Under five deaths per 1000 population
- `Polio`: Polio immunization coverage for 1 year olds
- `Diptheria`: Diphtheria tetanus toxoid and pertussis immunization coverage for 1 year olds
- `HIV.AIDS`: Deaths per 1000 live births from HIV/AIDS (0-4 year olds)
- `Income.composition.of.resources`: Human Development Index in terms of income composition of resources (from 0 to 1)
- `Schooling`: Number of Years of Schooling

We are interested in creating a regression model that predicts a country's life expectancy based on the other factors in the dataset.

a. Suppose we know that both adult mortality and income composition of resources are significant in a linear model that predicts life expectancy. We would like to know if adding development status signficantly improves the predictive ability of our model.

- i. [2 pts] Report the least squares regression lines for predicting life expectancy for both developed and developing countries with adult mortality and income composition of resources included.

```
lifeexp <- read.csv("../Data/lifeexp.csv")
mod <- lm(Life.expectancy ~ Status + Adult.Mortality + Income.composition.of.resources,lifeexp)
summary(mod)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Status + Adult.Mortality + Income.composition.of.resources,
##     data = lifeexp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2584  -1.5413  -0.1224   1.5715   7.6854
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     50.333822   2.127544  23.658  < 2e-16 ***
## StatusDeveloping                -0.429828   0.710997  -0.605    0.546
## Adult.Mortality                 -0.023829   0.003062  -7.781 7.11e-13 ***
## Income.composition.of.resources 36.627072   2.168108  16.894  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.893 on 167 degrees of freedom
## Multiple R-squared:  0.8676, Adjusted R-squared:  0.8652
## F-statistic: 364.8 on 3 and 167 DF,  p-value: < 2.2e-16
```

$$\hat{lifeexp}_{Developed} = 50.33 - 0.0238Mort + 36.63ICOR$$

$$\hat{lifeexp}_{Developing} = 49.90 - 0.0238Mort + 36.63ICOR$$

- ii. [1 pt] What is the expected difference in life expectancy between developed and developing countries for fixed levels of adult mortality and income composition of resources?

0.429 years

- iii. [4 pts] Perform a formal hypothesis test to determine if adding development status to our linear model significantly improves our model's predictive ability for life expectancy. Be sure to include all pieces of information needed to perform a hypothesis test. Assume $\alpha = 0.05$.

Reduced Model: $\hat{lifeexp} = \beta_0 + \beta_1 Mort + \beta_2 ICOR$ Full Model: $\hat{lifeexp} = \beta_0 + \beta_1 Mort + \beta_2 ICOR + \beta_3 I(Status = Developing)$

$$H_0 : \beta_3 = 0$$
$$H_a : \beta_3 \neq 0$$

```
mod_red <- lm(Life.expectancy ~ Adult.Mortality + Income.composition.of.resources,lifeexp)
mod_full <- lm(Life.expectancy ~ Status + Adult.Mortality + Income.composition.of.resources,lifeexp)
anova(mod_red,mod_full)
```

```
## Analysis of Variance Table
##
## Model 1: Life.expectancy ~ Adult.Mortality + Income.composition.of.resources
## Model 2: Life.expectancy ~ Status + Adult.Mortality + Income.composition.of.resources
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    168 1400.9
## 2    167 1397.8  1     3.059 0.3655 0.5463
```

$F = 0.3655$, $p = 0.5463$. We fail to reject H_0 and conclude that adding development status to a model that already includes adult mortality and income composition of resources does not significantly improve the predictive ability of life expectancy.

- iv. [5 pts] Perform a formal hypothesis test to determine if EITHER the linear relationship between adult mortality and life expectancy OR the linear relationship between income composition of resources and life expectancy is different for the two different development statuses. Be sure to include all pieces of information needed to perform a hypothesis test. Assume $\alpha = 0.05$. Reduced Model: $\hat{lifeexp} = \beta_0 + \beta_1 Mort + \beta_2 ICOR + \beta_3 I(Status = Developing)$ Full Model: $\hat{lifeexp} = \beta_0 + \beta_1 Mort + \beta_2 ICOR + \beta_3 I(Status = Developing) + \beta_4 Mort \times I(Status = Developing) + \beta_5 ICOR \times I(Status = Developing)$

$$H_0 : \beta_4 = \beta_5 = 0$$

$$H_a : \text{At least one } \beta \neq 0$$

```
mod_red <- lm(Life.expectancy ~ Status + Adult.Mortality + Income.composition.of.resources,lifeexp)
mod_full <- lm(Life.expectancy ~ Status + Adult.Mortality + Income.composition.of.resources + Status:Adu
anova(mod_red,mod_full)
```

```
## Analysis of Variance Table
##
## Model 1: Life.expectancy ~ Status + Adult.Mortality + Income.composition.of.resources
## Model 2: Life.expectancy ~ Status + Adult.Mortality + Income.composition.of.resources +
##     Status:Adult.Mortality + Status:Income.composition.of.resources
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    167 1397.8
## 2    165 1377.9  2    19.869 1.1896 0.3069
```

$F = 1.190$, $p = 0.3069$. We fail to reject H_0 and conclude that adding either an interaction term for income composition of resources and development status or an interaction term for adult mortality and development status to a model that already includes development status, adult mortality, and income composition of resources does not significantly improve the predictive ability of life expectancy.

b. We now would like to determine the "best" overall prediction model for life expectancy using all of the other variables in our dataset as predictors. Assume no polynomial or interaction terms are significant in our model.

- i. [5 pts] Use backward selection with AIC as your decision criteria to eliminate variables from your model. Be sure to state which variables were eliminated at each step.

```
mod_max <- lm(Life.expectancy ~ .,lifeexp)
mod <- step(mod_max)
```

```
## Start:  AIC=347.22
## Life.expectancy ~ Status + Adult.Mortality + infant.deaths +
##     Measles + BMI + under.five.deaths + Polio + Diphtheria +
##     HIV.AIDS + Income.composition.of.resources + Schooling
##
##                                   Df Sum of Sq    RSS    AIC
## - Schooling                        1      0.64 1132.8 345.32
## - Measles                          1      1.27 1133.4 345.41
## - BMI                              1      3.09 1135.2 345.69
## - Polio                            1      7.60 1139.7 346.36
## - Status                           1     10.50 1142.6 346.80
## <none>                                       1132.1 347.22
## - Diphtheria                       1     22.30 1154.4 348.56
## - infant.deaths                    1     25.63 1157.7 349.05
## - under.five.deaths                1     30.77 1162.9 349.81
## - HIV.AIDS                         1     75.46 1207.6 356.25
## - Adult.Mortality                  1    202.41 1334.5 373.35
## - Income.composition.of.resources  1    445.75 1577.8 401.99
##
```

```
## Step:  AIC=345.32
## Life.expectancy ~ Status + Adult.Mortality + infant.deaths +
##     Measles + BMI + under.five.deaths + Polio + Diphtheria +
##     HIV.AIDS + Income.composition.of.resources
##
##                                  Df Sum of Sq    RSS    AIC
## - Measles                         1      1.10 1133.8 343.48
## - BMI                             1      2.88 1135.6 343.75
## - Polio                           1      7.37 1140.1 344.43
## - Status                          1     11.70 1144.5 345.07
## <none>                                      1132.8 345.32
## - Diphtheria                      1     22.94 1155.7 346.75
## - infant.deaths                   1     25.29 1158.0 347.09
## - under.five.deaths               1     30.51 1163.3 347.86
## - HIV.AIDS                        1     74.94 1207.7 354.27
## - Adult.Mortality                 1    204.12 1336.9 371.65
## - Income.composition.of.resources  1   1296.47 2429.2 473.78
##
## Step:  AIC=343.48
## Life.expectancy ~ Status + Adult.Mortality + infant.deaths +
##     BMI + under.five.deaths + Polio + Diphtheria + HIV.AIDS +
##     Income.composition.of.resources
##
##                                  Df Sum of Sq    RSS    AIC
## - BMI                             1      2.87 1136.7 341.91
## - Polio                           1      7.37 1141.2 342.59
## - Status                          1     11.91 1145.8 343.27
## <none>                                      1133.8 343.48
## - Diphtheria                      1     22.35 1156.2 344.82
## - infant.deaths                   1     29.12 1163.0 345.82
## - under.five.deaths               1     32.34 1166.2 346.29
## - HIV.AIDS                        1     75.15 1209.0 352.46
## - Adult.Mortality                 1    203.52 1337.4 369.71
## - Income.composition.of.resources  1   1295.42 2429.3 471.78
##
## Step:  AIC=341.91
## Life.expectancy ~ Status + Adult.Mortality + infant.deaths +
##     under.five.deaths + Polio + Diphtheria + HIV.AIDS + Income.composition.of.resources
##
##                                  Df Sum of Sq    RSS    AIC
## - Polio                           1      7.97 1144.7 341.11
## <none>                                      1136.7 341.91
## - Status                          1     13.39 1150.1 341.92
## - Diphtheria                      1     23.31 1160.0 343.39
## - infant.deaths                   1     29.69 1166.4 344.32
## - under.five.deaths               1     32.68 1169.4 344.76
## - HIV.AIDS                        1     73.49 1210.2 350.63
## - Adult.Mortality                 1    209.16 1345.9 368.80
## - Income.composition.of.resources  1   1637.40 2774.1 492.48
##
## Step:  AIC=341.11
## Life.expectancy ~ Status + Adult.Mortality + infant.deaths +
##     under.five.deaths + Diphtheria + HIV.AIDS + Income.composition.of.resources
##
```

```
##                                   Df Sum of Sq    RSS    AIC
## - Status                           1     12.70 1157.4 340.99
## <none>                                         1144.7 341.11
## - infant.deaths                    1     29.92 1174.6 343.52
## - under.five.deaths                1     33.00 1177.7 343.97
## - Diphtheria                       1     55.77 1200.5 347.24
## - HIV.AIDS                         1     75.43 1220.1 350.02
## - Adult.Mortality                  1    214.34 1359.0 368.46
## - Income.composition.of.resources  1   1720.51 2865.2 496.00
##
## Step:  AIC=340.99
## Life.expectancy ~ Adult.Mortality + infant.deaths + under.five.deaths +
##     Diphtheria + HIV.AIDS + Income.composition.of.resources
##
##                                   Df Sum of Sq    RSS    AIC
## <none>                                        1157.4 340.99
## - infant.deaths                    1     26.98 1184.4 342.93
## - under.five.deaths                1     29.89 1187.3 343.35
## - Diphtheria                       1     56.29 1213.7 347.12
## - HIV.AIDS                         1     70.14 1227.5 349.06
## - Adult.Mortality                  1    218.33 1375.7 368.55
## - Income.composition.of.resources  1   2357.80 3515.2 528.96
```

The model removed schooling, then measles, then bmi, then polio, then status to arrive at our final model.

- ii. [3 pts] Write your final regression model for predicting life expectancy.

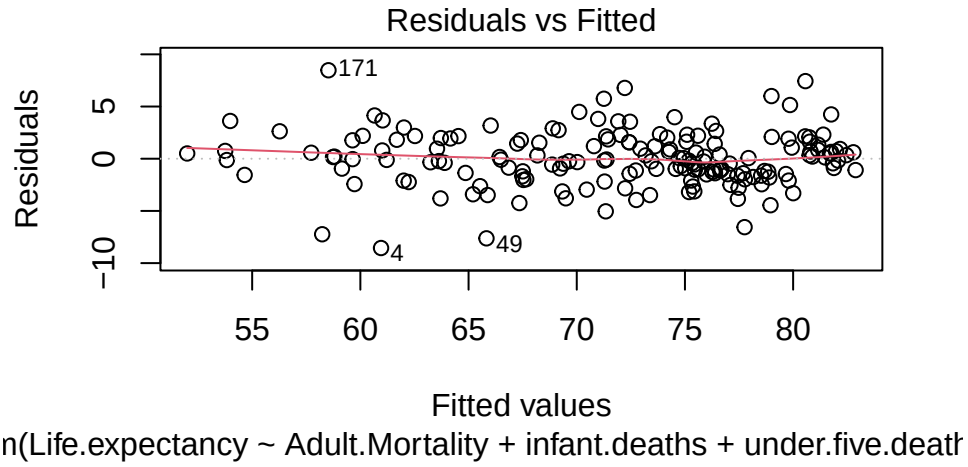
```
summary(mod)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + infant.deaths +
##     under.five.deaths + Diphtheria + HIV.AIDS + Income.composition.of.resources,
##     data = lifeexp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5555 -1.4542 -0.0938  1.6278  8.4750
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     48.694210   1.645845  29.586  < 2e-16 ***
## Adult.Mortality                 -0.017831   0.003206  -5.562 1.06e-07 ***
## infant.deaths                    0.044730   0.022878   1.955  0.05227 .
## under.five.deaths               -0.036806   0.017885  -2.058  0.04118 *
## Diphtheria                       0.030184   0.010687   2.824  0.00533 **
## HIV.AIDS                        -0.638743   0.202612  -3.153  0.00192 **
## Income.composition.of.resources 34.220999   1.872211  18.278  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.657 on 164 degrees of freedom
## Multiple R-squared:  0.8904, Adjusted R-squared:  0.8864
## F-statistic:   222 on 6 and 164 DF,  p-value: < 2.2e-16
```

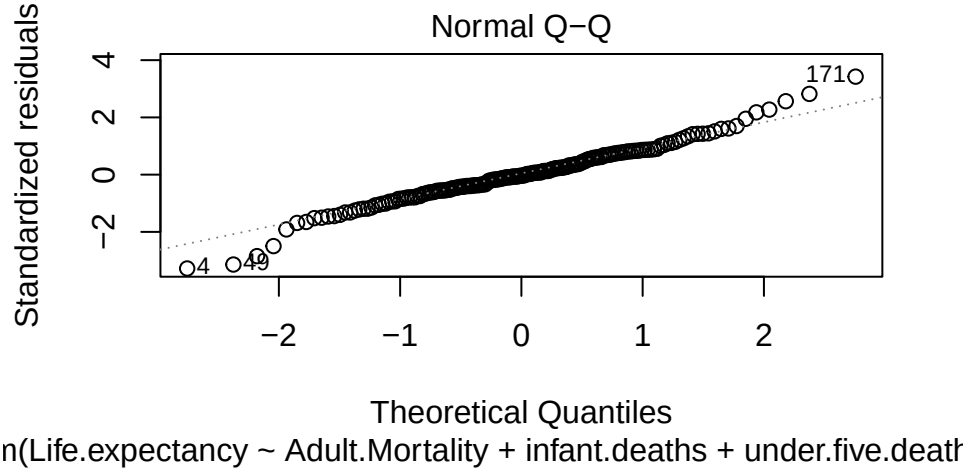
$\hat{LifeExp} = 48.69 - 0.0178Mort + 0.0447Infant.Deaths - 0.0368Under.Five.Deaths + 0.0302Diph - 0.6387HIV + 34.22ICOR$

- iii. [2 pts] Determine if the assumptions of homoscedasticity and normally distributed residuals for the final regression model are violated.

```
plot(mod,1)
```




```
plot(mod,2)
```



The points on the residual plot look fairly evenly spread for all fitted values of the model and have an average close to zero, so the assumption of homoscedasticity does not appear to be violated. The points within -2 and 2 appear close to the 45-degree line of the QQ plot with some slight deviation in the tails. This is not too signficant, so I would say the assumption of normally distributed residuals is not violated.