HW3

Adhrit Srivastav, eid: Ams22362

9/16/2021

Problem 1 in Canvas

```
airplane = c(24, 31, 32, 39, 47, 47, 35, 76, 95, 85)
helicopter = c(30, 30, 33, 38, 58, 58, 48, 75, 85, 55)

modelM = lm(airplane ~ helicopter)
summary(modelM)
```

```
##
## Call:
## lm(formula = airplane ~ helicopter)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.6551  -9.8500  -0.5449   3.7176  29.3068
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.4632    12.4957  -0.597  0.56685
## helicopter    1.1483     0.2311   4.969  0.00109 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.13 on 8 degrees of freedom
## Multiple R-squared:  0.7553, Adjusted R-squared:  0.7247
## F-statistic: 24.69 on 1 and 8 DF,  p-value: 0.001094
```

```
anova(modelM)
```

```
## Analysis of Variance Table
##
## Response: airplane
##            Df Sum Sq Mean Sq F value   Pr(>F)
## helicopter  1 4259.0  4259.0  24.692 0.001094 **
## Residuals   8 1379.9   172.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# t stat for first test
1.1483/0.2311
```

```
## [1] 4.968845
```

```r
# t stat for second test
q = (1.1483-1)/0.2311
p_value=2*pt(q, df=8, lower.tail=FALSE)


# part D, no intercept
modelM0 = lm(airplane ~ helicopter + 0)
summary(modelM0)
```

```
##
## Call:
## lm(formula = airplane ~ helicopter + 0)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.8700 -10.6744  -0.9788   0.4200  29.0031
##
## Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
## helicopter  1.01813    0.07401   13.76 2.39e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.66 on 9 degrees of freedom
## Multiple R-squared:  0.9546, Adjusted R-squared:  0.9496
## F-statistic: 189.3 on 1 and 9 DF,  p-value: 2.386e-07
```

```r
anova(modelM0)
```

```
## Analysis of Variance Table
##
## Response: airplane
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## helicopter  1 30309.6 30309.6  189.25 2.386e-07 ***
## Residuals   9  1441.4   160.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# t stat for second test
q0 = (1.01813-1)/0.07401
p_value0=2*pt(q0, df=9, lower.tail=FALSE)
```

a) Helicopters are explanatory and airplanes are response. They wanted to use helicopter counts to check airplane counts.

b) $\widehat{Airplane}$ = -7.4632 + *helicopter**1.1483; For every additional manatee count from a helicopter, the number of manatees counted from an airplane increased by approximately on the average 1.1483. The

intercept means that if zero manatees are counted from a helicopter, there will be -7.4632, or zero, manatees counted from an airplane. The coefficient of determination (R^2) is .7553 which means .7553 is the proportion of the variation in manatees counted from an airplane explained by the regression model.

c) In test one where H_0: $B_1 = 0$, the t value is 4.969 with a pval of .001094 with degrees of freedom of 8. Because the p-val is less than .05 we reject null and say there is significant association between airplane and helicopter manatee counts.

In the second test where H_0: $B_1 = 1$, the t value is .6417 with a p-val is .539 with degrees of freedom of 9. We fail to reject null because the p-value is greater than .05 thus there is not significant evidence that there is not a one-for-one increase. Relevant test is the second one because we expect the one-for-one increase and that is what the researchers are testing for.

d) $\widehat{Airplane} = helicopter$ * 1.01813

In test one where H_0: $B_1 = 0$, the t value is 13.76 with a pval of 2.386e-07 with degrees of freedom of 9. Because the p-val is less than .05 we reject null and say there is significant association between airplane and helicopter manatee counts with the relation between the two counts going through the origin.

In the second test where H_0: $B_1 = 1$, the t value is .244966 with a p-val is .81197. We fail to reject null because the p-value is greater than .05 thus there is not significant evidence that there is not a one-for-one increase. Relevant test is the second one because we expect the one-for-one increase and that is what the researchers are testing for.

3.2 Breakfast Cereals

a) $\widehat{Calories} = 116.6$

b) The residual is 6.6 which means the predicted caloric value is 6.6 greater than the actual caloric value.

3.4 Breakfast cereals: understanding estimates

The correlation coefficient and other representative values of correlation are not given, so we can't say with confidence whether or not the amount of sugar has a weaker relationship with the number of calories than the amount of fiber. The formula coefficients alone are not enough to suggest that the amount of sugar has a weaker relationship with the number of calories than the amount of fiber.

3.6 Breakfast cereals: interpreting estimates

For every additional gram of fiber per serving, the calories per serving goes down by approximately 3.7 calories after adjusting/controlling for grams of sugar per serving.

3.9 Body Measurements

a) I do believe that BodyFat and Waist are positively correlated. While someone could innately have a broader bone structure that could result in a larger Waist size, I think a primary proponent would be BodyFat. Humans tend to gain fat in their abdomen area first, thus a larger Waist size would likely have a larger BodyFat percentage.

b) Since waist size is fixed, taller people might have a higher metabolism and thus a lower BodyFat. I can see there being a negative correlation between Height and BodyFat.

c) Similar to part B, there would likely be a negative correlation between Height and BodyFat percentage. Considering this, I believe the coefficient on Height would likely be negative in a multiple regression to predict BodyFat based on Height and Waist.

3.13 Models for well water

a) $Arsenic = B_0 + B_1$*Year + B_2*Miles + B_3*Years*Miles + ϵ

b) $Lead = B_0 + B_1$*Year + B_2*IClean + B_3*Years*IClean + ϵ

c) $Titanium = B_0 + B_1$*Miles + B_2*$Miles^2$ + ϵ

d) $Sulfide = B_0 + B_1$*Year + B_2*Miles + B_3*Depth + B_4*Years*Miles + B_5*Years*Depth B_6*Miles*Depth + ϵ

3.16 Predicting faculty salaries

a) $Salary = B_0 + B_1$*Age*Seniority + B_2*Age*Pub + B_3*Age*IGender + B_4*Seniority*Pub + B_5*Seniority*IGender + B_6*IGender*Pub + ϵ

b) I believe Age and Seniority will be correlated because the older a faculty member is the more likely they are to have more years of experience. The younger a faculty member is the less likely they are to have large numbers of years of experience.

c) I believe Seniority and Pub will be correlated because the more years of experience a faculty member has, the more likely they are to have created publications. Naturally, when a faculty member has less years of experience, they are less likely to have large numbers of publications because publications take many years to develop.

d) I do think the dean will be unhappy because that means there is a significant difference in the amount men are getting paid over women. This would indicate a gender-based pay gap which would not be acceptable.

3.22 Breakfast cereal:

a) The R^2 is .5439. This means that .5439 is the proportion of the variance in Calories explained by the regression model.

b) The regression standard error is 15.413.

c) The F-ratio for testing the null hypothesis is 19.677.

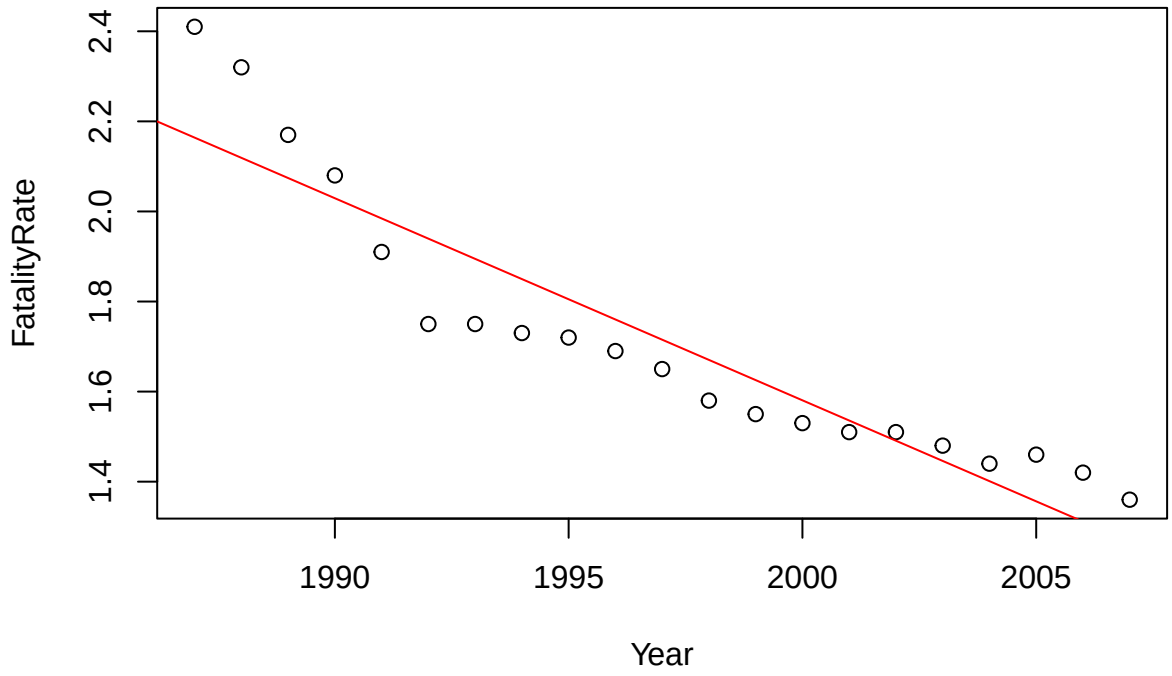
d) The small p-value of .000002 indicates that at least one of the variables of sugar or fiber is useful for predicting calories in breakfast cereals. The p-value is small enough that we can reject the null hypothesis.

3.32 Driving fatalities and speed limits

```
speed = read.csv("C:/Users/adhri/OneDrive/Documents/R/App_Reg_and_Time_Series/chpt3/datasets/Speed.csv")
attach(speed)

# a
plot(FatalityRate ~ Year)
```

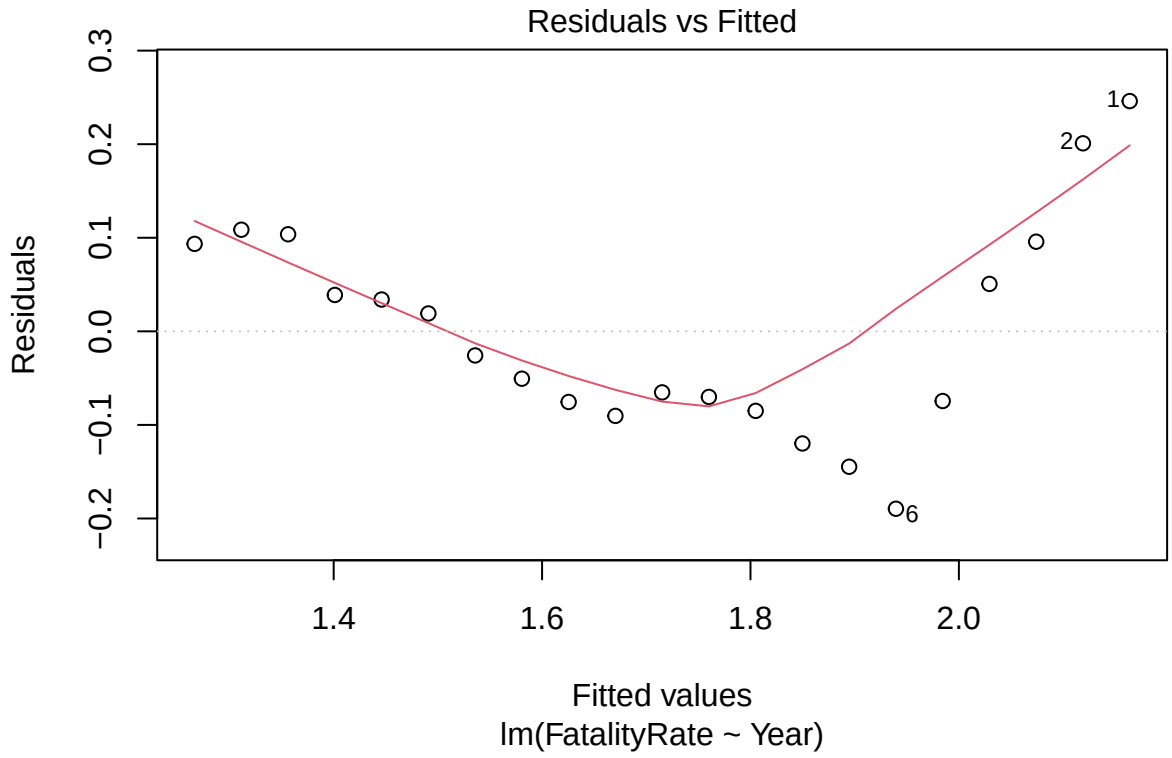
```r
modelS = lm(FatalityRate ~ Year)
abline(modelS, col = 'red')
```

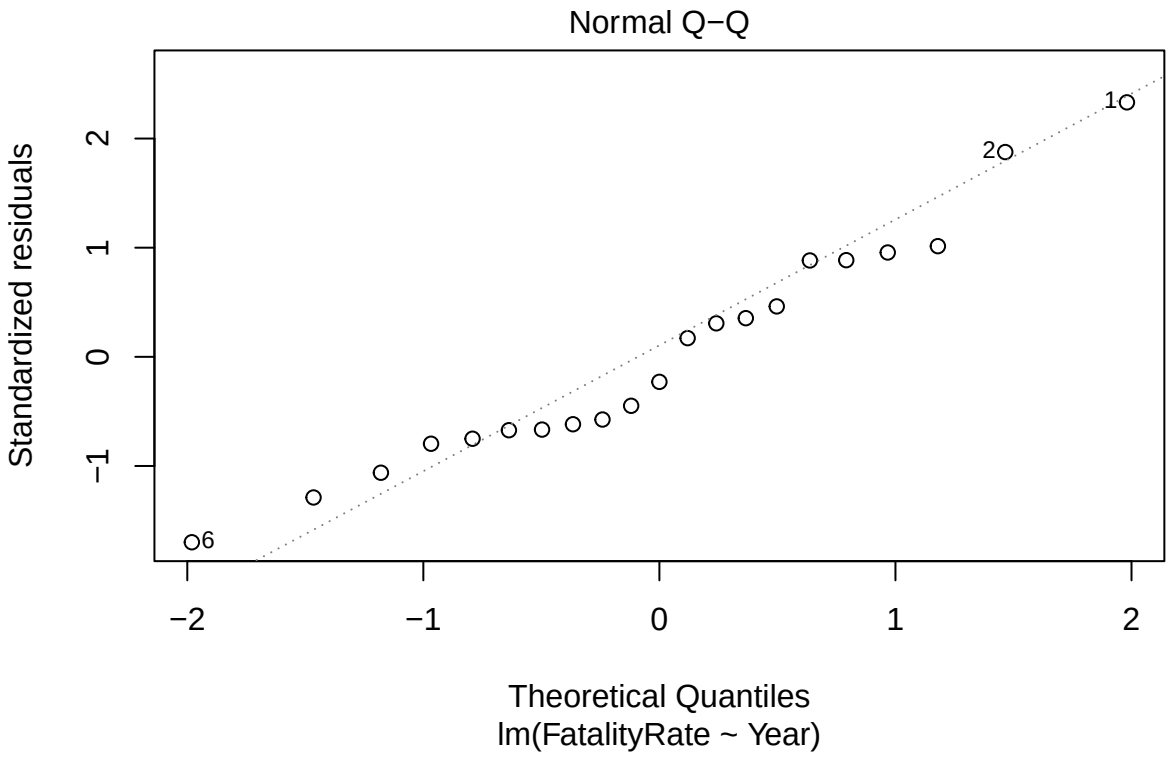


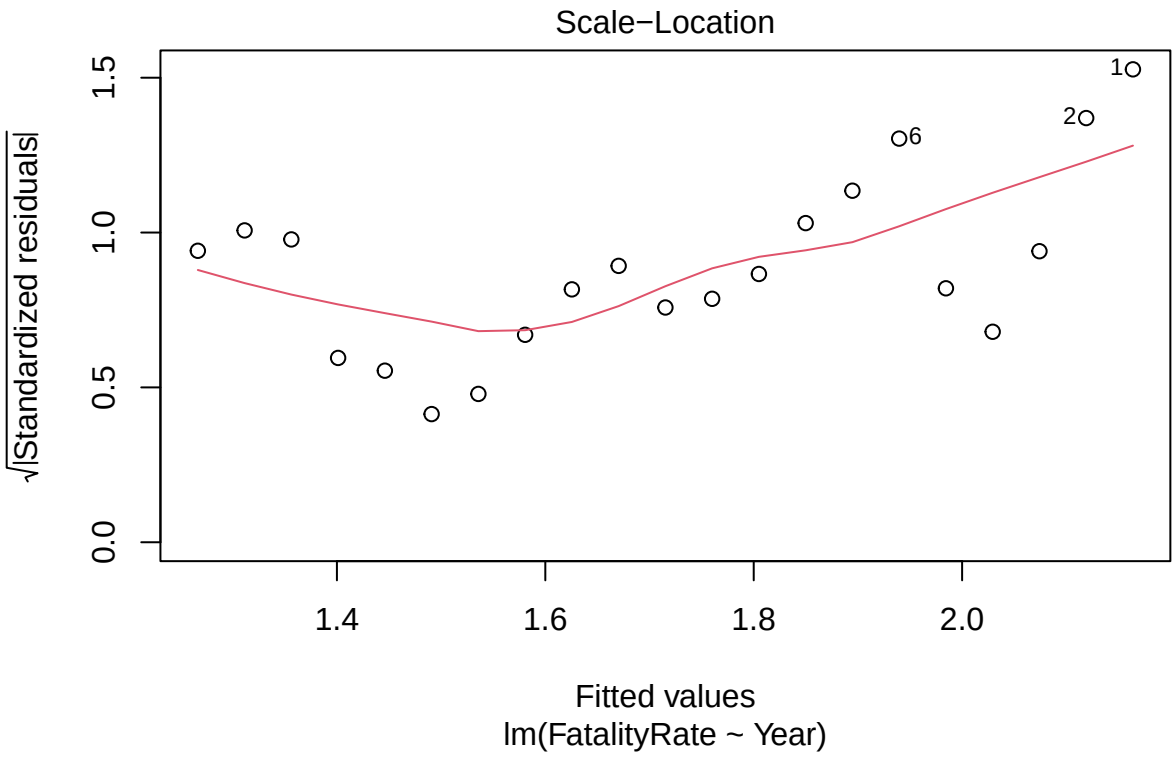

```r
summary(modelS)
```

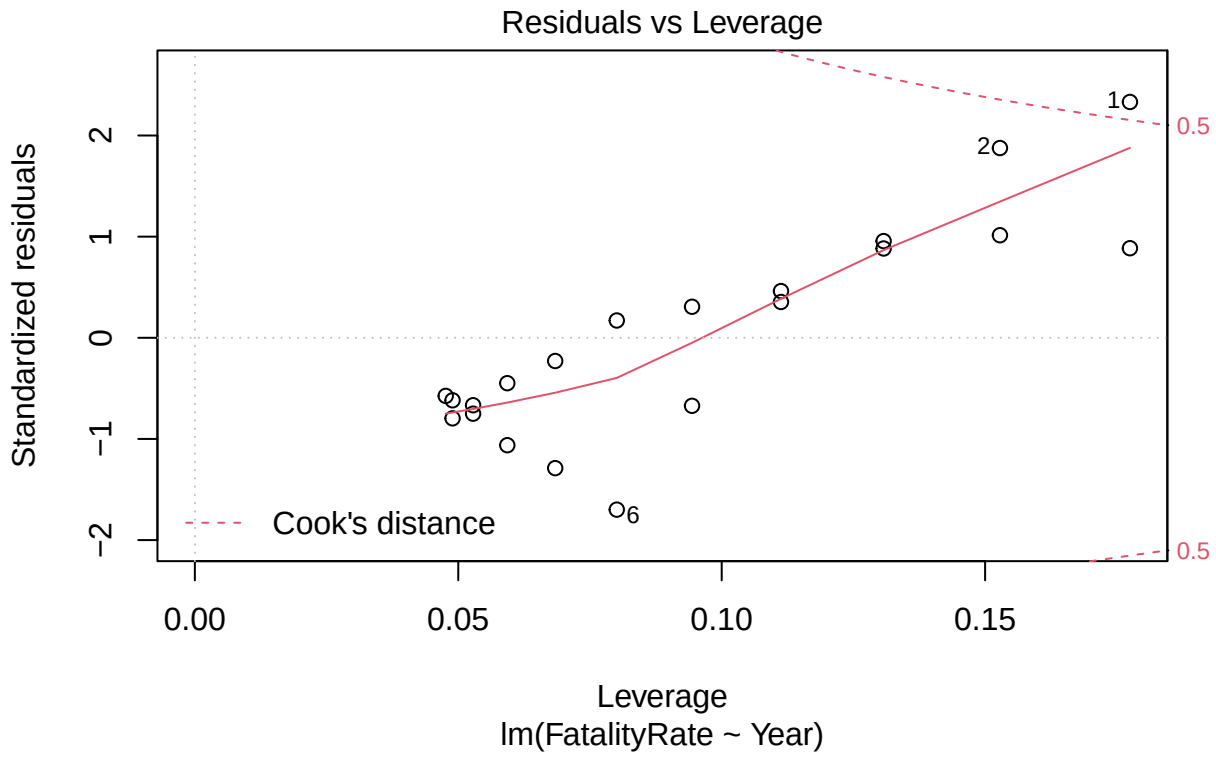
```
##
## Call:
## lm(formula = FatalityRate ~ Year)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18959 -0.07550 -0.02576  0.09346  0.24606
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 91.320887   8.374227    10.9 1.28e-09 ***
## Year        -0.044870   0.004193   -10.7 1.75e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1164 on 19 degrees of freedom
## Multiple R-squared:  0.8577, Adjusted R-squared:  0.8502
## F-statistic: 114.5 on 1 and 19 DF,  p-value: 1.75e-09
```

```r
# b
plot(modelS)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(FatalityRate ~ Year)
1
2
6
−2
−1
0
1
2

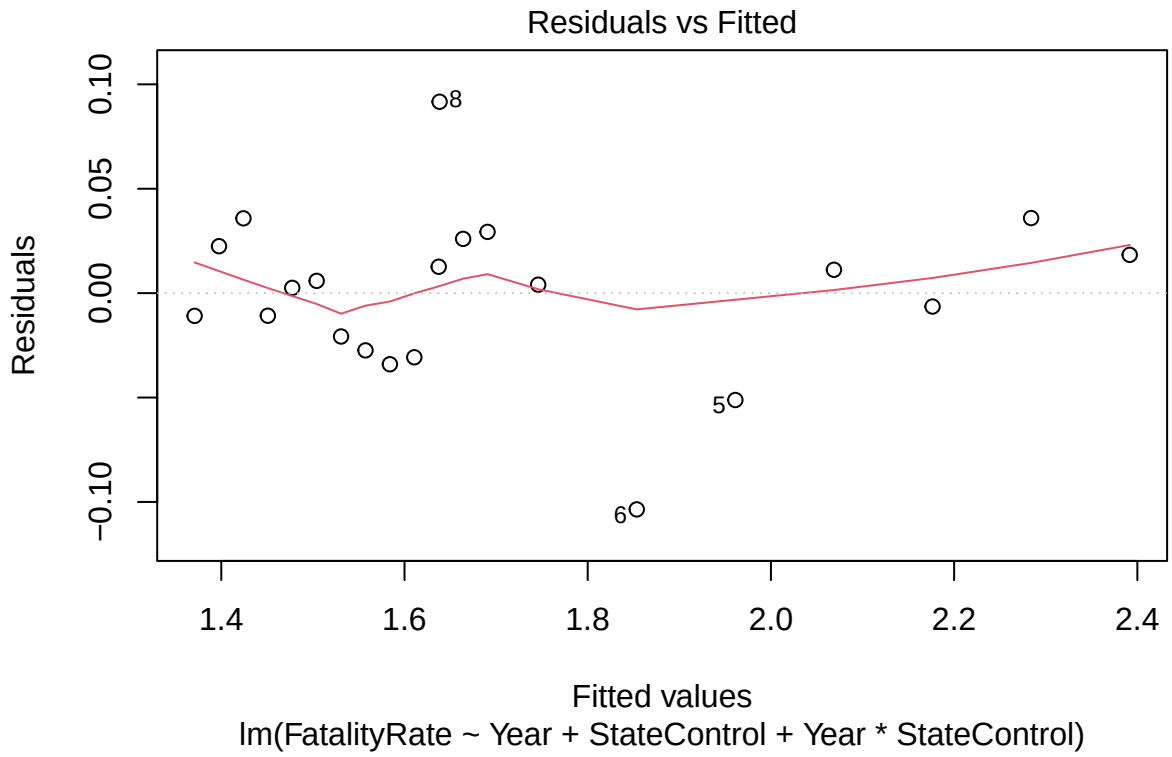

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
1.4
1.6
1.8
2.0
Fitted values
lm(FatalityRate ~ Year)
1
2
6

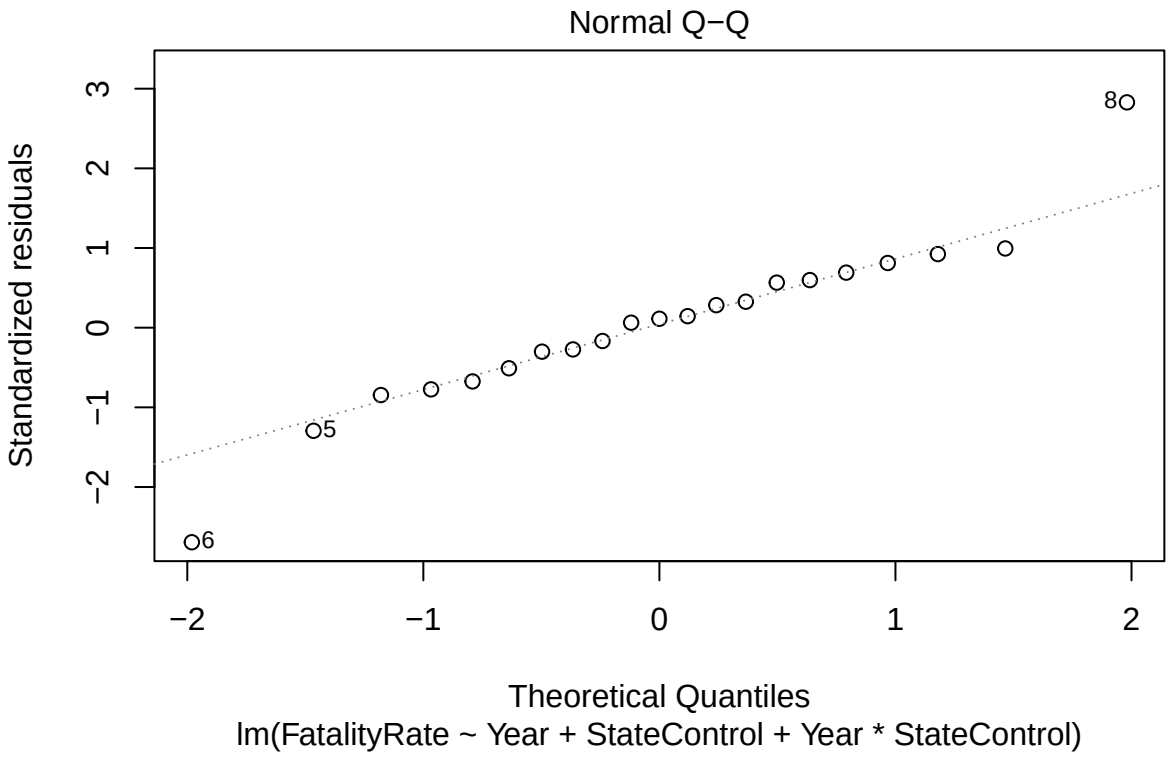


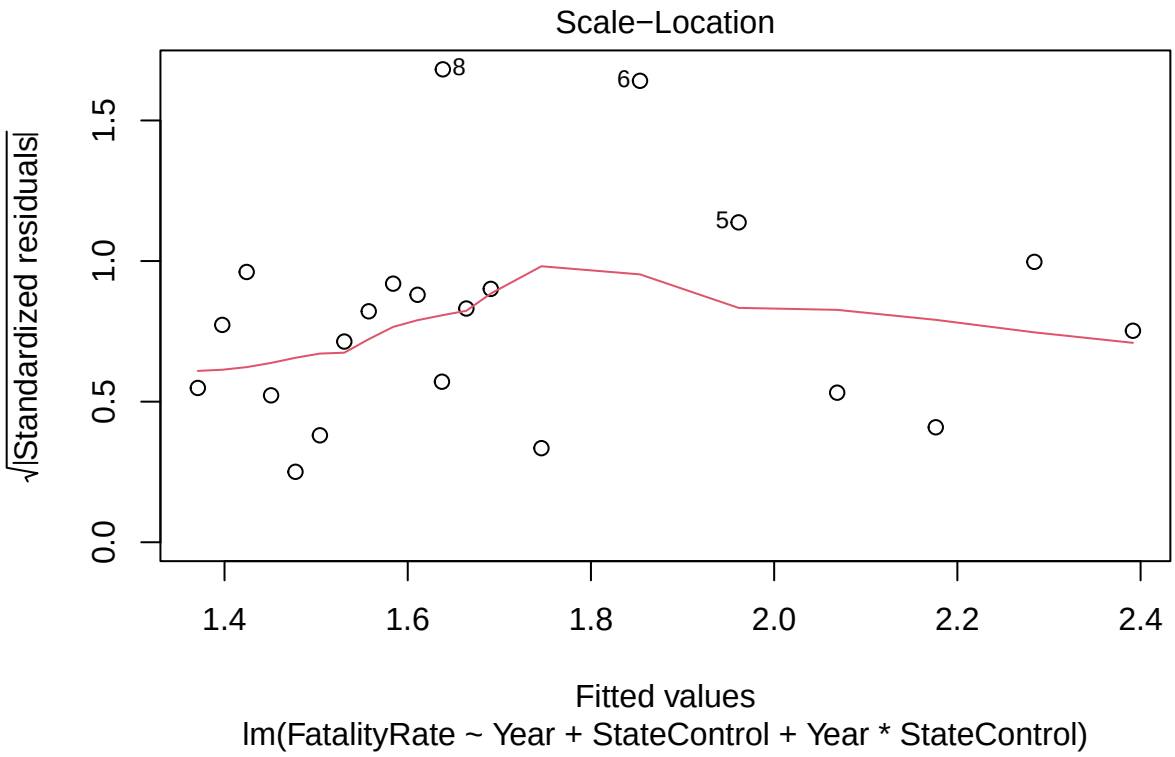

```
# c
modelMS = lm(FatalityRate ~ Year + StateControl + Year*StateControl)
summary(modelMS)
```

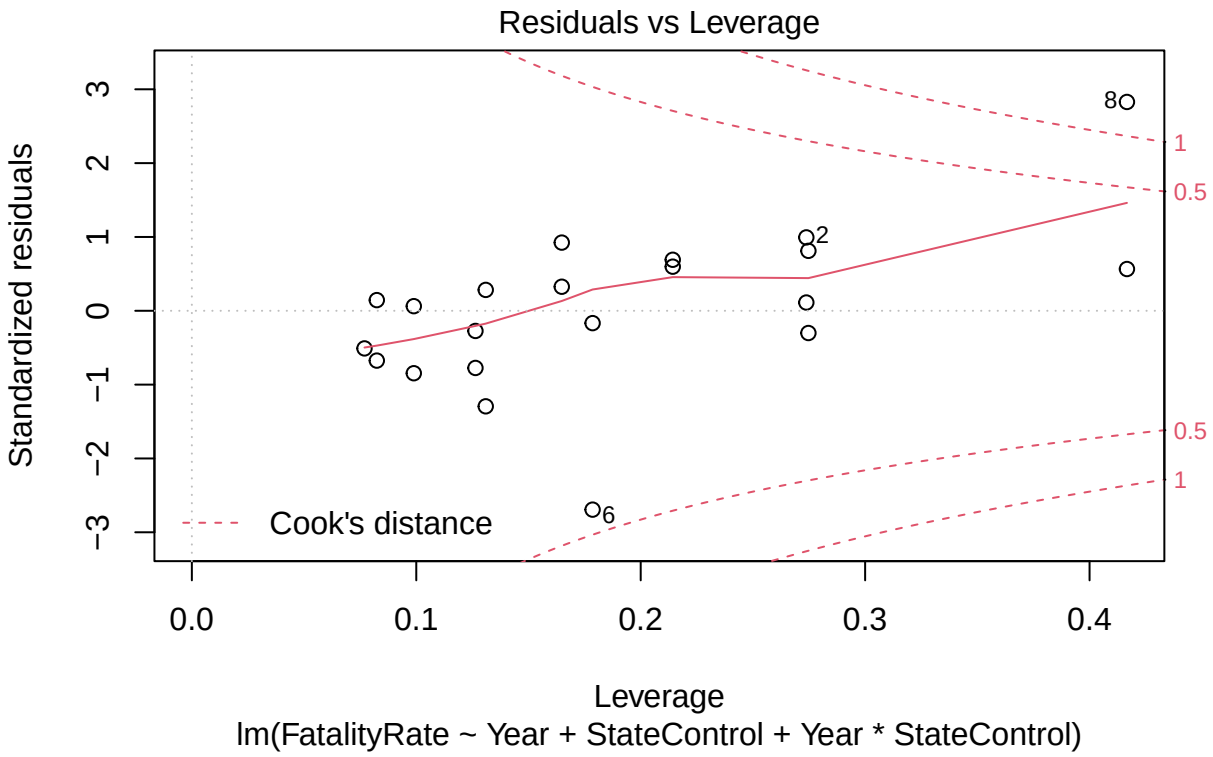
```
##
## Call:
## lm(formula = FatalityRate ~ Year + StateControl + Year * StateControl)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.103571 -0.020769  0.004048  0.022473  0.091667
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.162e+02  1.303e+01   16.59 6.19e-12 ***
## Year              -1.076e-01  6.548e-03  -16.44 7.19e-12 ***
## StateControl      -1.614e+02  1.447e+01  -11.15 3.07e-09 ***
## Year:StateControl  8.097e-02  7.264e-03   11.15 3.08e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04243 on 17 degrees of freedom
## Multiple R-squared:  0.9831, Adjusted R-squared:  0.9801
## F-statistic:   329 on 3 and 17 DF,  p-value: 2.998e-15
```

```
plot(modelMS)
```




Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
8
5
6
−2
−1
0
1
2
Theoretical Quantiles
lm(FatalityRate ~ Year + StateControl + Year * StateControl)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
8
6
5
1.4
1.6
1.8
2.0
2.2
2.4
Fitted values
lm(FatalityRate ~ Year + StateControl + Year * StateControl)



a) The slope is -0.044870.

b) The residual plot has a curvature to it which means it fails the linearity test. The residual plot rebounded.

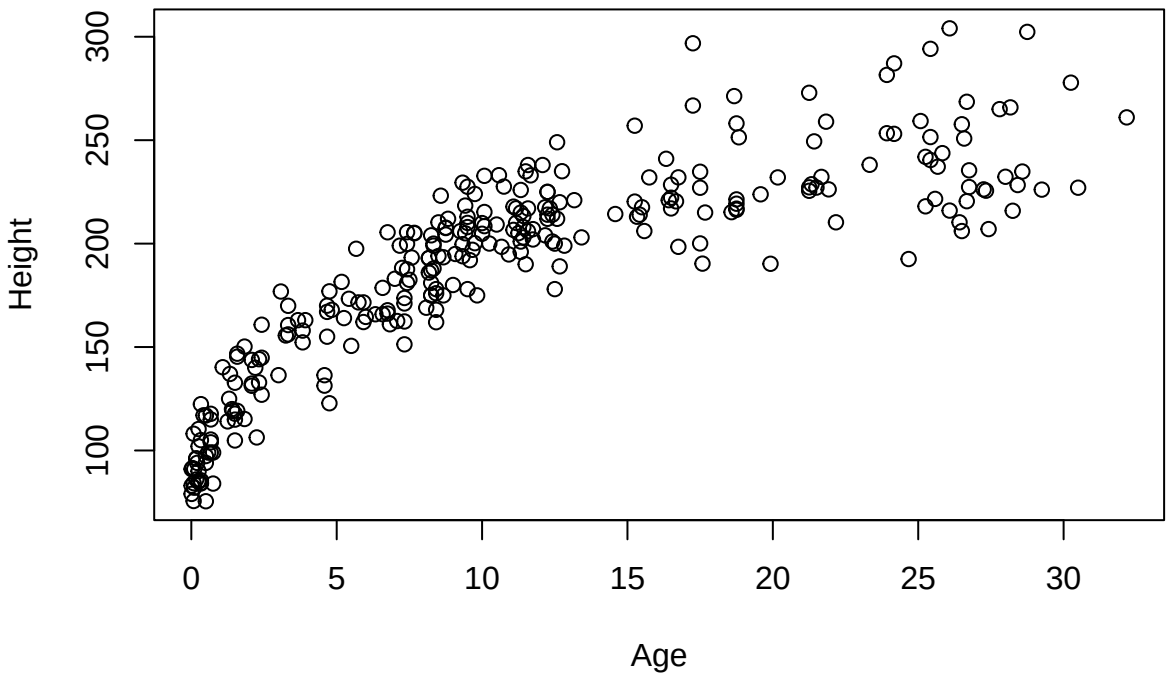
c) I believe there is a significant change in relationship between fatality rate and year starting in 1995 because the coefficient between fatality rate and year change significantly especially when considering the fact that the data is based on fatalities per 100 million vehicle-miles of travel.

d) Pre 1995: $\widehat{FatalityRate}$ = 2.162e+02 + $\widehat{Year}$*-1.076e-01
Post 1995: $\widehat{FatalityRate}$ = 54.8 + $\widehat{Year}$*0.02663

3.36 Elephants, gender

```
elph = read.csv("C:/Users/adhri/OneDrive/Documents/R/App_Reg_and_Time_Series/chpt3/datasets/ElephantsMF
attach(elph)

plot(Height ~ Age)
```




```
modelE = lm(Height~Age, method = "qr")
#abline(modelE, col = 'red')
summary(modelE)
```

```
##
## Call:
## lm(formula = Height ~ Age, method = "qr")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.141 -18.041   2.681  19.969  77.532
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 132.4221     2.7027   49.00   <2e-16 ***
## Age           5.0369     0.1958   25.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.84 on 286 degrees of freedom
## Multiple R-squared:  0.6983, Adjusted R-squared:  0.6973
## F-statistic:   662 on 1 and 286 DF,  p-value: < 2.2e-16
```

```
# creating quadratic model
Age2 = Age^2
```

```
modelQE = lm(Height ~ Age + Age2)
summary(modelQE)     # we can see the R^2 value for the quadratic model is .85 whereas basic linear mod
```

```
##
## Call:
## lm(formula = Height ~ Age + Age2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.910 -13.337  -1.226  11.900  66.968
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 102.48332    2.54514   40.27   <2e-16 ***
## Age          12.56560    0.45204   27.80   <2e-16 ***
## Age2         -0.27628    0.01582  -17.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.38 on 285 degrees of freedom
## Multiple R-squared:  0.8543, Adjusted R-squared:  0.8533
## F-statistic: 835.5 on 2 and 285 DF,  p-value: < 2.2e-16
```

```
#lines(hourValues, happinessPredict, col='red')

# c
102.48332 + 10*12.56560 + 10^2*-0.27628
```

```
## [1] 200.5113
```

a) The plot has a positively curved, exponential pattern/nature.

b) $\widehat{Height} = 102.48332 + \widehat{Age}*12.56560 + \widehat{Age}^2*-0.27628$

c) The predicted height of a 10-year-old elephant is 200.5113 cm.