Exam 1

Adhrit Srivastav, eid: Ams22362

9/21/2021

Problem 1

```
data1 = read.csv('C:/Users/adhri/OneDrive/Documents/R/App_Reg_and_Time_Series/exam1/ques1data.csv', file
attach(data1)

# a
mod = lm(cost ~ cargotype)
summary(mod)
```

```
##
## Call:
## lm(formula = cost ~ cargotype)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -2.20  -1.80  -1.00   1.05   4.24
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.260      1.075   3.032   0.0104 *
## cargotypeF     9.740      1.521   6.405 3.38e-05 ***
## cargotypeS     5.440      1.521   3.577   0.0038 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.404 on 12 degrees of freedom
## Multiple R-squared:  0.7745, Adjusted R-squared:  0.7369
## F-statistic: 20.61 on 2 and 12 DF,  p-value: 0.0001315
```

```
# b
data1$F = ifelse(data1$cargotype == "F", 1, 0)
data1$S = ifelse(data1$cargotype == "S", 1, 0)
data1$D = ifelse(data1$cargotype == "D", 1, 0)
attach(data1)
```

```
## The following objects are masked from data1 (pos = 3):
##
##     cargotype, cost
```

```
## The following object is masked from package:base:
##
##     F
```

```
data1
```

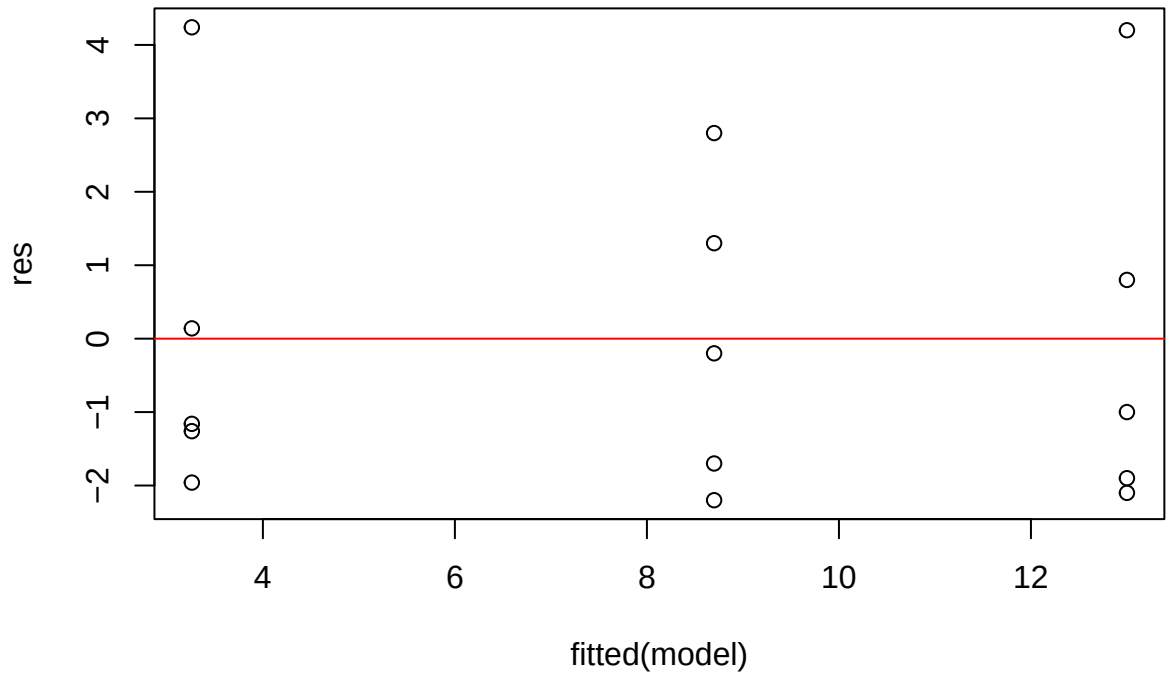
```
##    cost cargotype F S D
## 1  17.2         F 1 0 0
## 2  11.1         F 1 0 0
## 3  12.0         F 1 0 0
## 4  10.9         F 1 0 0
## 5  13.8         F 1 0 0
## 6   6.5         S 0 1 0
## 7  10.0         S 0 1 0
## 8  11.5         S 0 1 0
## 9   7.0         S 0 1 0
## 10  8.5         S 0 1 0
## 11  2.1         D 0 0 1
## 12  1.3         D 0 0 1
## 13  3.4         D 0 0 1
## 14  7.5         D 0 0 1
## 15  2.0         D 0 0 1
```

```
# plot(cost ~ F + S)

model = lm(cost ~ F + S)
summary(model)
```

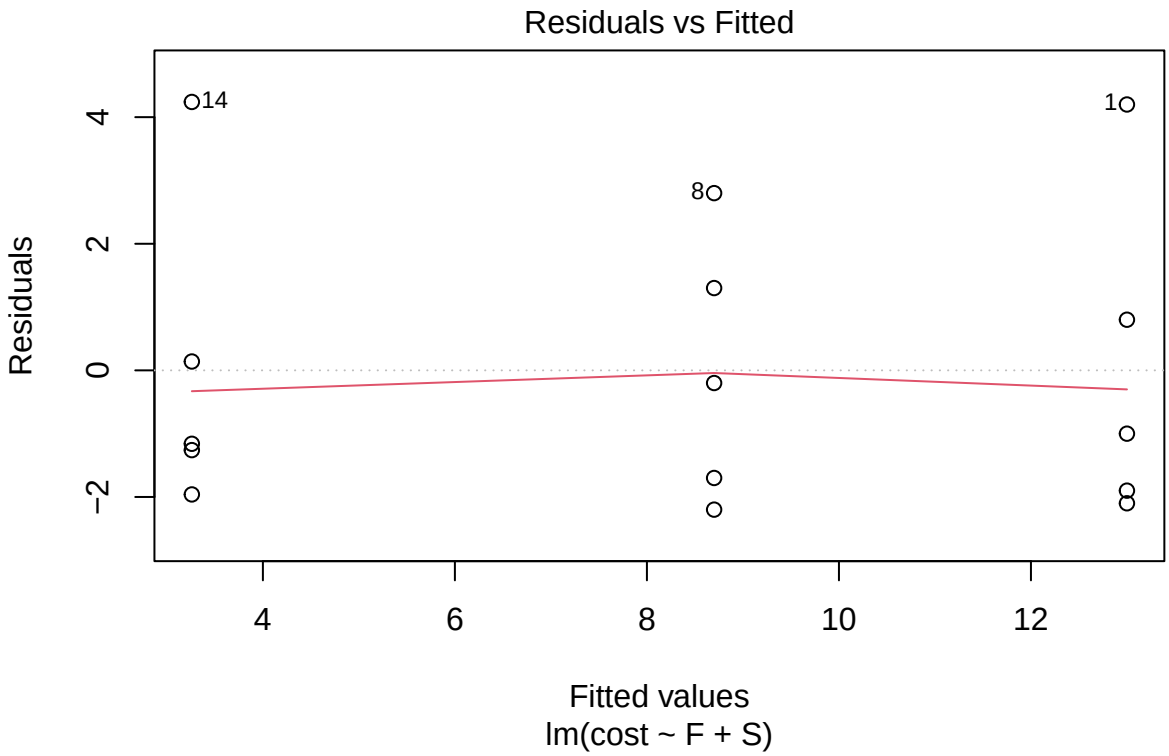
```
##
## Call:
## lm(formula = cost ~ F + S)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -2.20  -1.80  -1.00   1.05   4.24
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.260      1.075   3.032   0.0104 *
## F              9.740      1.521   6.405 3.38e-05 ***
## S              5.440      1.521   3.577   0.0038 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.404 on 12 degrees of freedom
## Multiple R-squared:  0.7745, Adjusted R-squared:  0.7369
## F-statistic: 20.61 on 2 and 12 DF,  p-value: 0.0001315
```

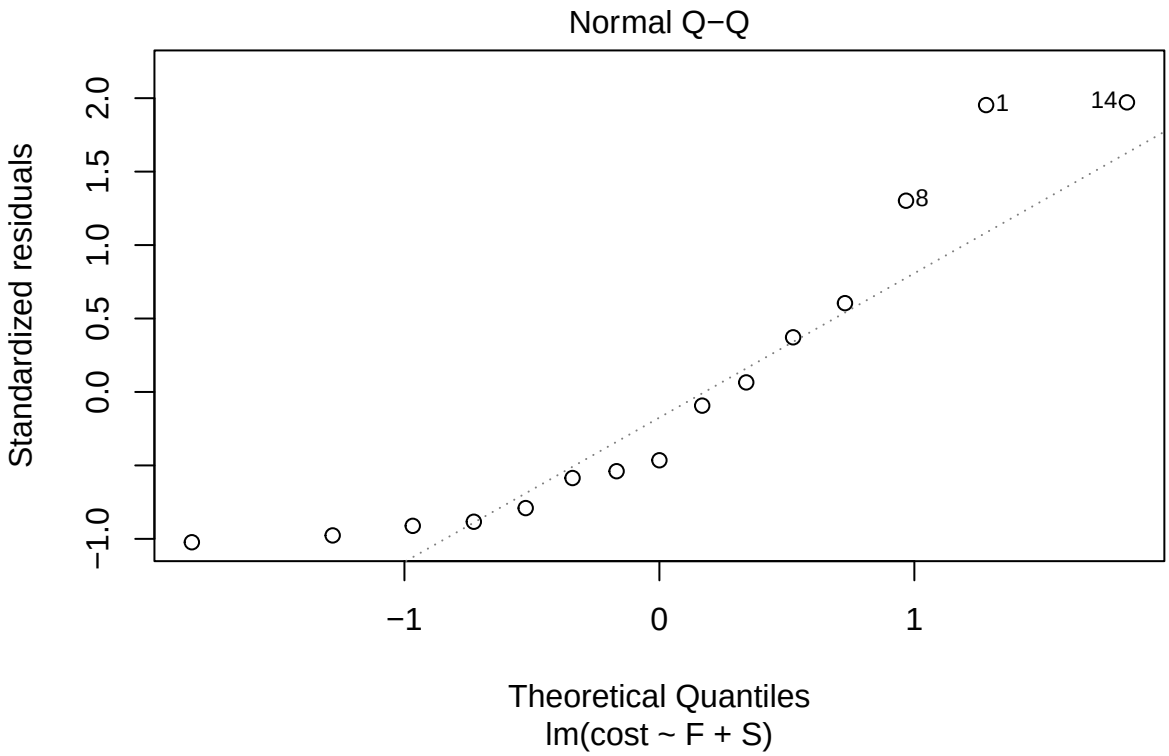
```
# e
# residual plot

#re = resid(mod)      # same as model
```

```r
#plot(fitted(mod), re)
#abline(h = 0, col = 'red')

res = resid(model)
plot(fitted(model), res)
abline(h = 0, col = 'red')
```

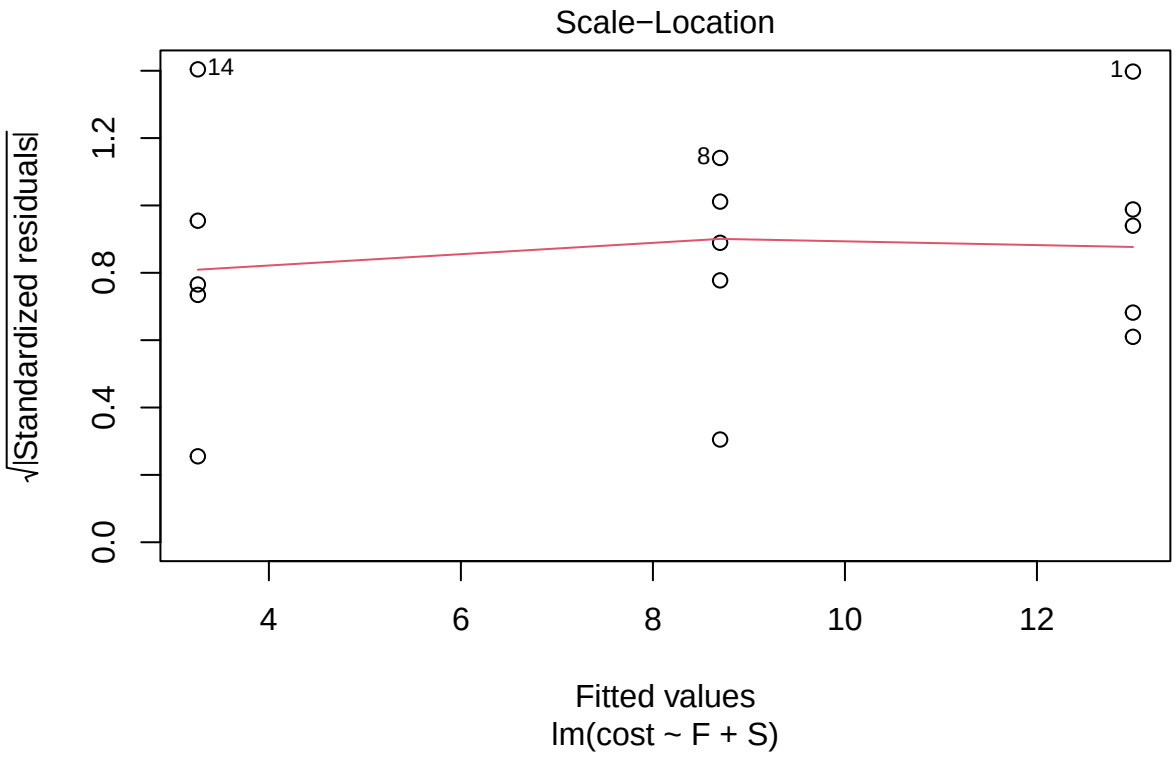



```r
#plot(mod)      # same as model
plot(model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(cost ~ F + S)
14
1
8
4
2
0
−2
4
6
8
10
12




```
## hat values (leverages) are all = 0.2
##  and there are no factor predictors; no plot no. 5
```


Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
14
8
1
4
6
8
10
12
Fitted values
lm(cost ~ F + S)

a) $\hat{cost}$ = 3.26 + 9.74*F + 5.44*S

b) I encoded/coded the dummies by making the code look for "F", "S", and "D" values in the cargotype column in the data. If the code found "F", it would create a new column in the data1 dataframe labeled "F" and put a 1 for every occurance of F in the cargotype column, and a 0 for all other values. I used similar code for the "s" and "D" values in the cargotype column and thus created three additional columns with the variables in cargotype successfully encoded/coded into dummies. The relevant output is the printed data1 dataframe above.

c) For every additional Fragile cargo type, the cost increases approximately on the average by 9.74 after adjusting/controlling for semi-fragile cargo type. For every additional Semi-fragile cargo type, the cost increases approximately on the average by 5.44 after adjusting/controlling for Fragile cargo type.

d) Using the model, the predicted cost for the first observation is 13. The first observation in the dataset is 17.2. Thus the residual is 4.2.

e) The residual plot shows equal variance for the three cargo types as there is no thickening in the plot. The residuals are randomly spread out and have independence. There is also no curvature in the Scale-Location plot indicating that it passes the linearity test. The Q-Q plot also appears to be straight thus the normality test is also passed.

Problem 2

```
data2 = read.csv('C:/Users/adhri/OneDrive/Documents/R/App_Reg_and_Time_Series/exam1/ques2data.csv', fil
attach(data2)
```

```r
# a
model1 = lm(Y ~ X1)
model2 = lm(Y ~ X2)
model12 = lm(Y ~ X1 + X2)

# b
summary(model1)
```

```
##
## Call:
## lm(formula = Y ~ X1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.16910 -0.67912 -0.00326  0.64412  1.12993
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.988755   1.266891   9.463  3.4e-07 ***
## X1           0.003747   0.416083   0.009    0.993
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8324 on 13 degrees of freedom
## Multiple R-squared:  6.24e-06,   Adjusted R-squared:  -0.07692
## F-statistic: 8.112e-05 on 1 and 13 DF,  p-value: 0.993
```

```r
summary(model2)
```

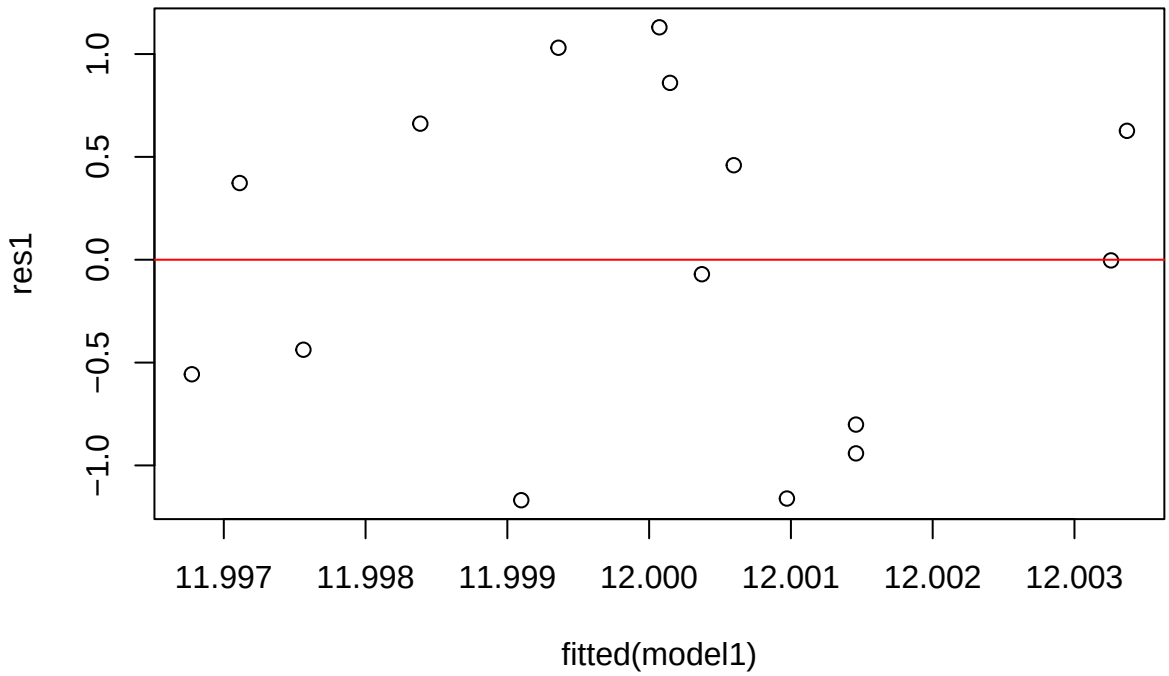
```
##
## Call:
## lm(formula = Y ~ X2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.08999 -0.63345  0.00023  0.61458  1.04033
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.6319     0.8109  13.111 7.18e-09 ***
## X2            0.1955     0.1125   1.737    0.106
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7499 on 13 degrees of freedom
## Multiple R-squared:  0.1884, Adjusted R-squared:  0.126
## F-statistic: 3.018 on 1 and 13 DF,  p-value: 0.106
```

```r
summary(model12)
```

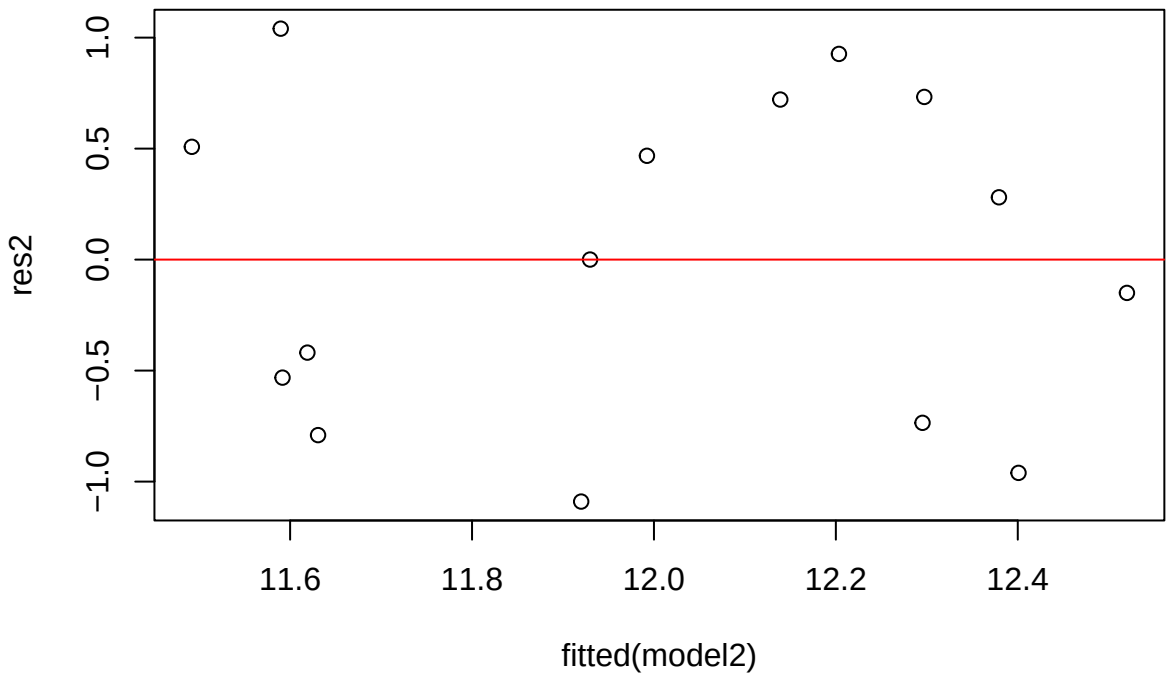
```
##
```

```
## Call:
## lm(formula = Y ~ X1 + X2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.013632 -0.009451 -0.002279  0.008630  0.016325
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.515414   0.061142  -73.85   <2e-16 ***
## X1           3.097008   0.012274  252.31   <2e-16 ***
## X2           1.031859   0.003684  280.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01072 on 12 degrees of freedom
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9998
## F-statistic: 3.922e+04 on 2 and 12 DF,  p-value: < 2.2e-16
```

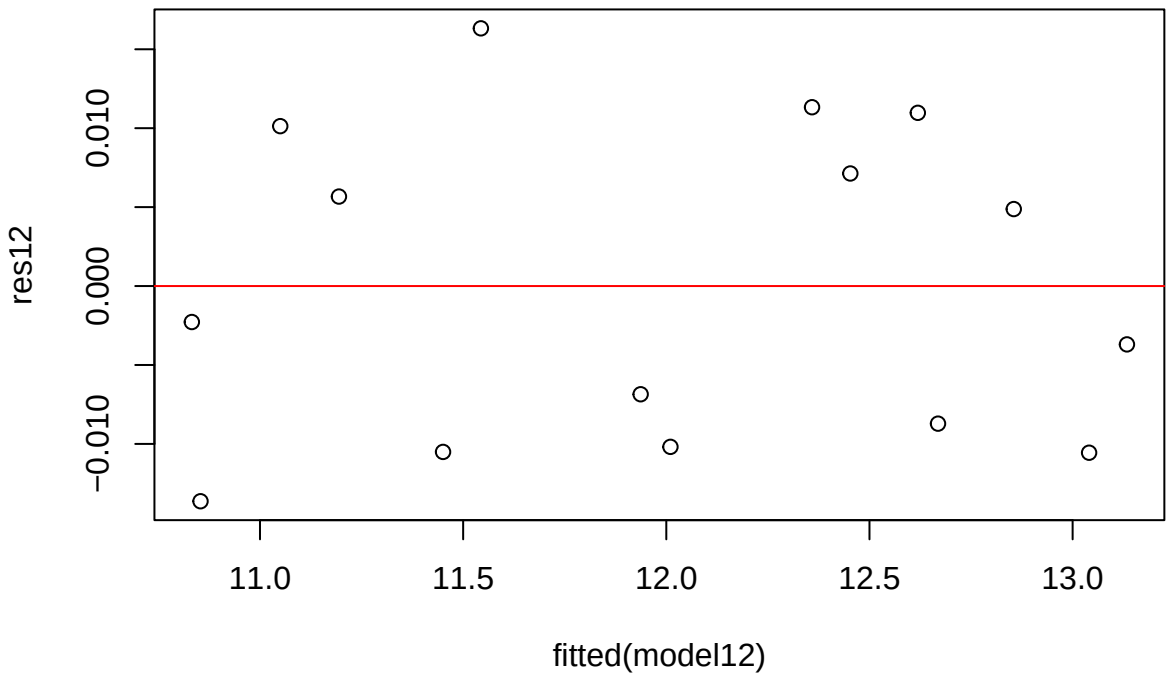
```r
# c
# residual plot
res1 = resid(model1)
plot(fitted(model1), res1)
abline(h = 0, col = 'red')
```



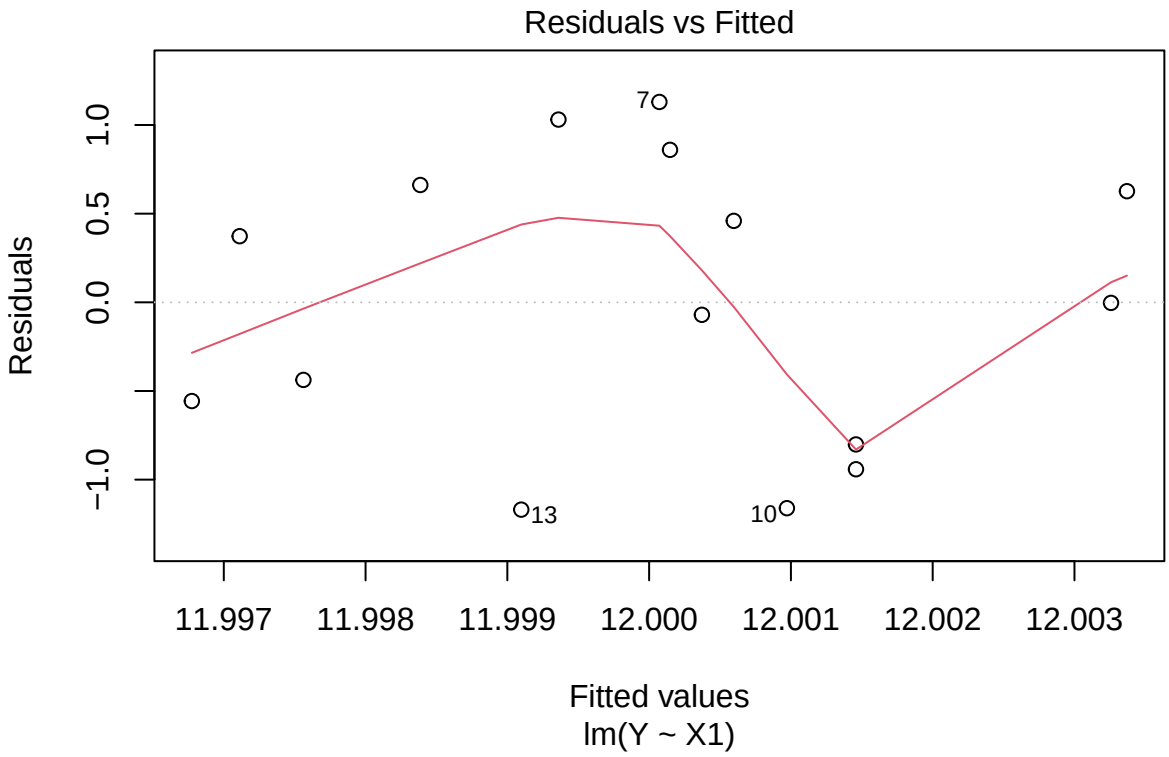
```r
res2 = resid(model2)
plot(fitted(model2), res2)
abline(h = 0, col = 'red')
```

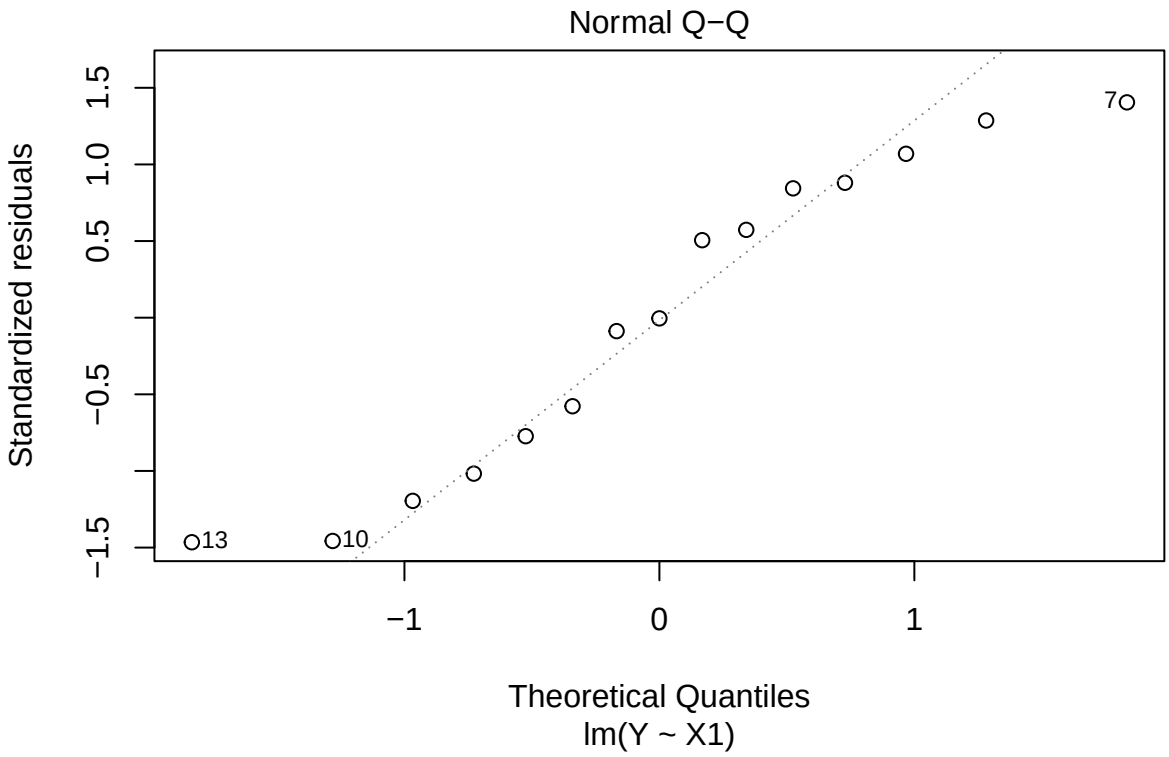


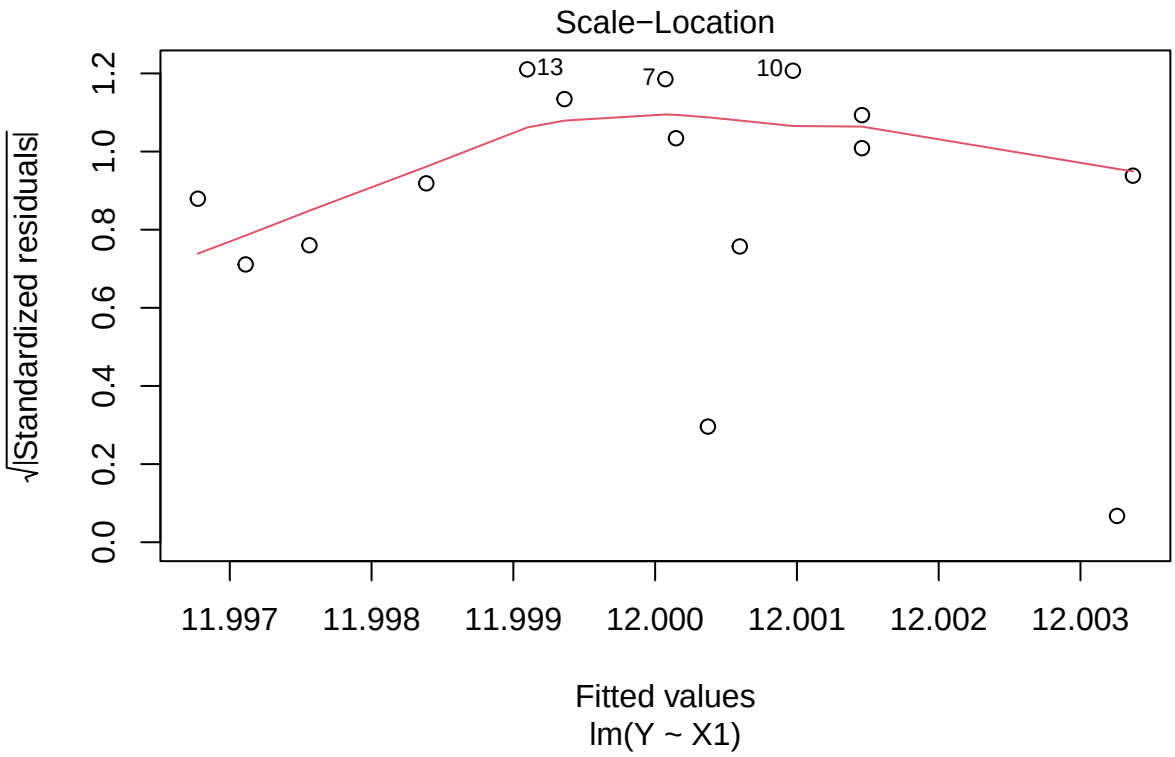

```r
res12 = resid(model12)
plot(fitted(model12), res12)
abline(h = 0, col = 'red')
```

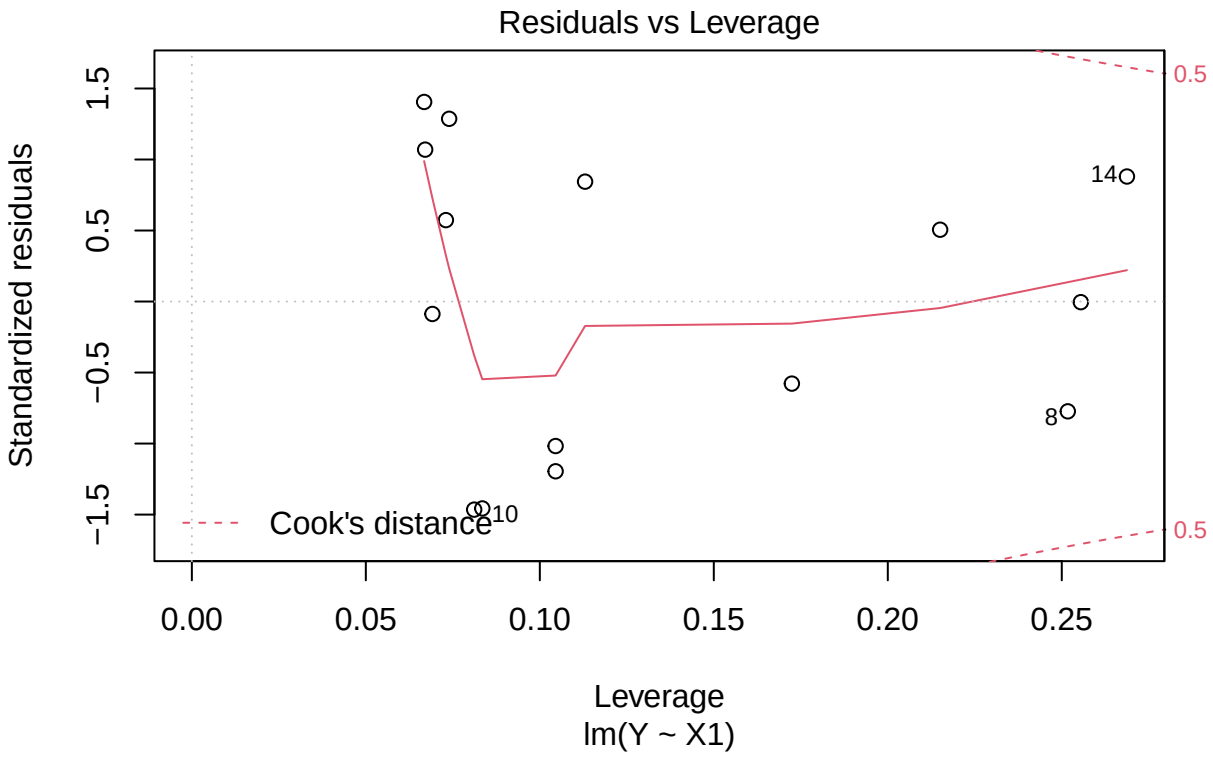



```
plot(model1)
```

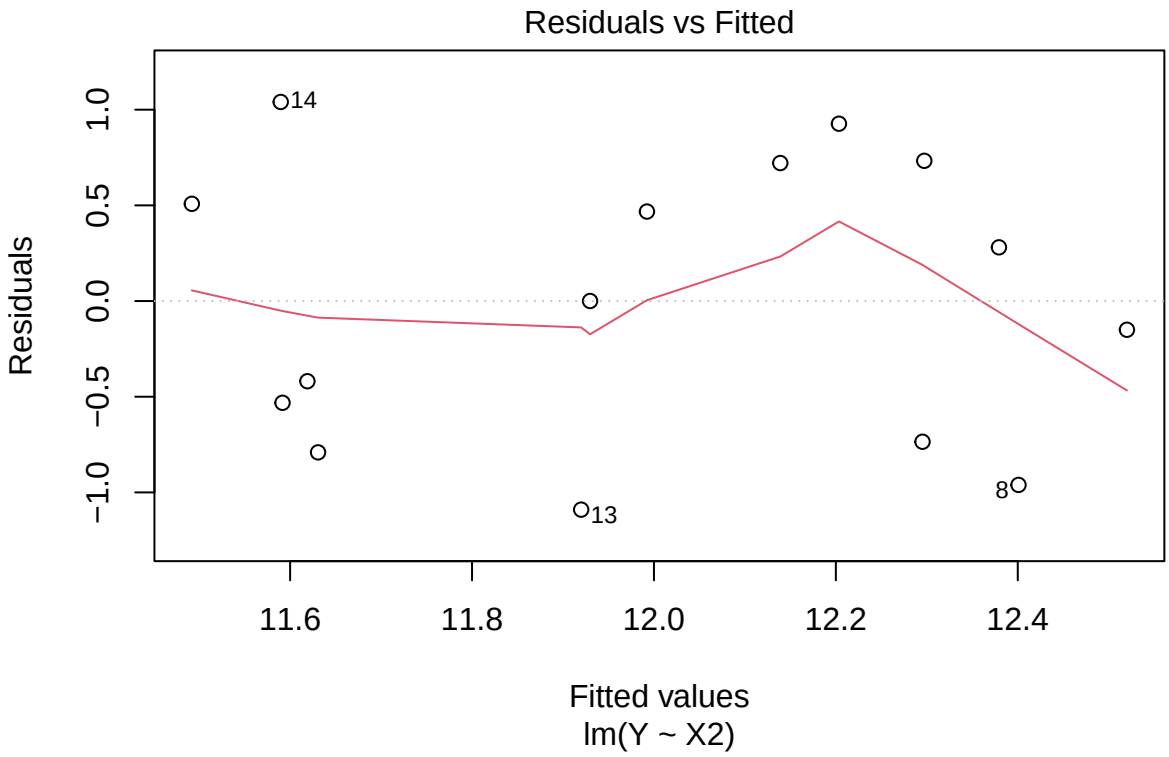

Residuals vs Fitted
Residuals
1.0
0.5
0.0
−1.0
7
13
10
11.997
11.998
11.999
12.000
12.001
12.002
12.003
Fitted values
lm(Y ~ X1)

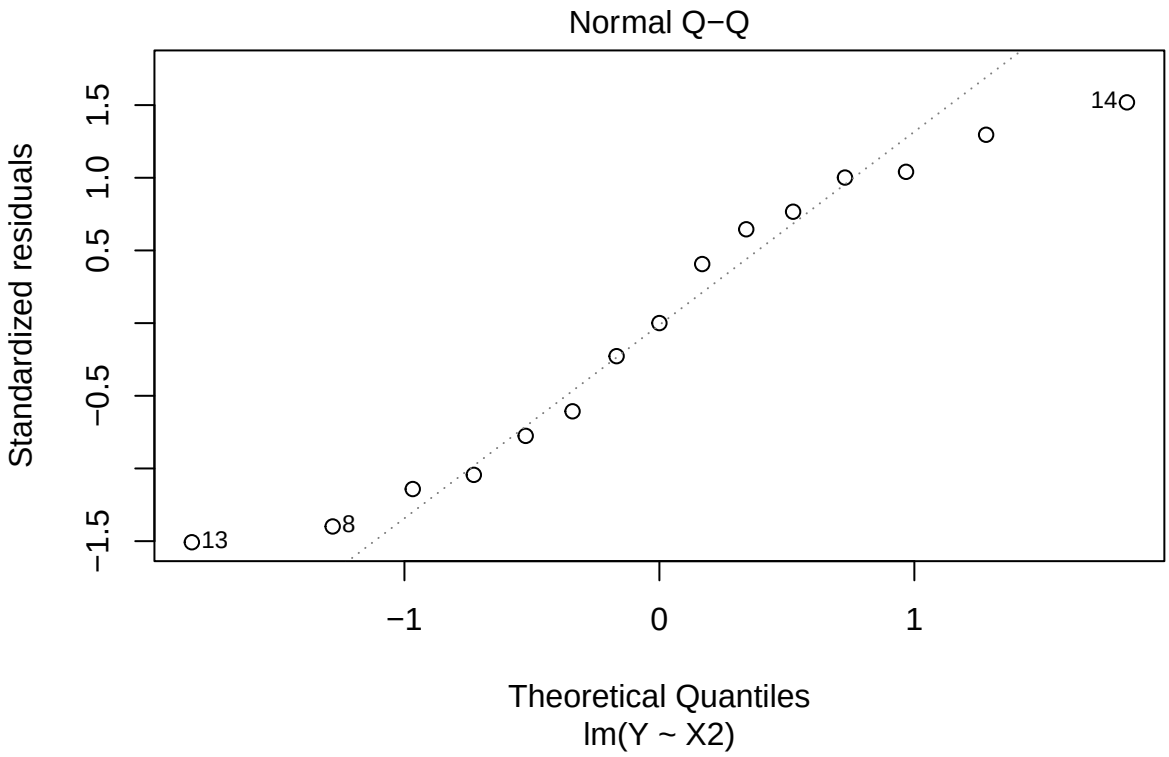

Normal Q−Q
Standardized residuals
−1.5
−0.5
0.5
1.0
1.5
13
10
7
−1
0
1
Theoretical Quantiles
lm(Y ~ X1)

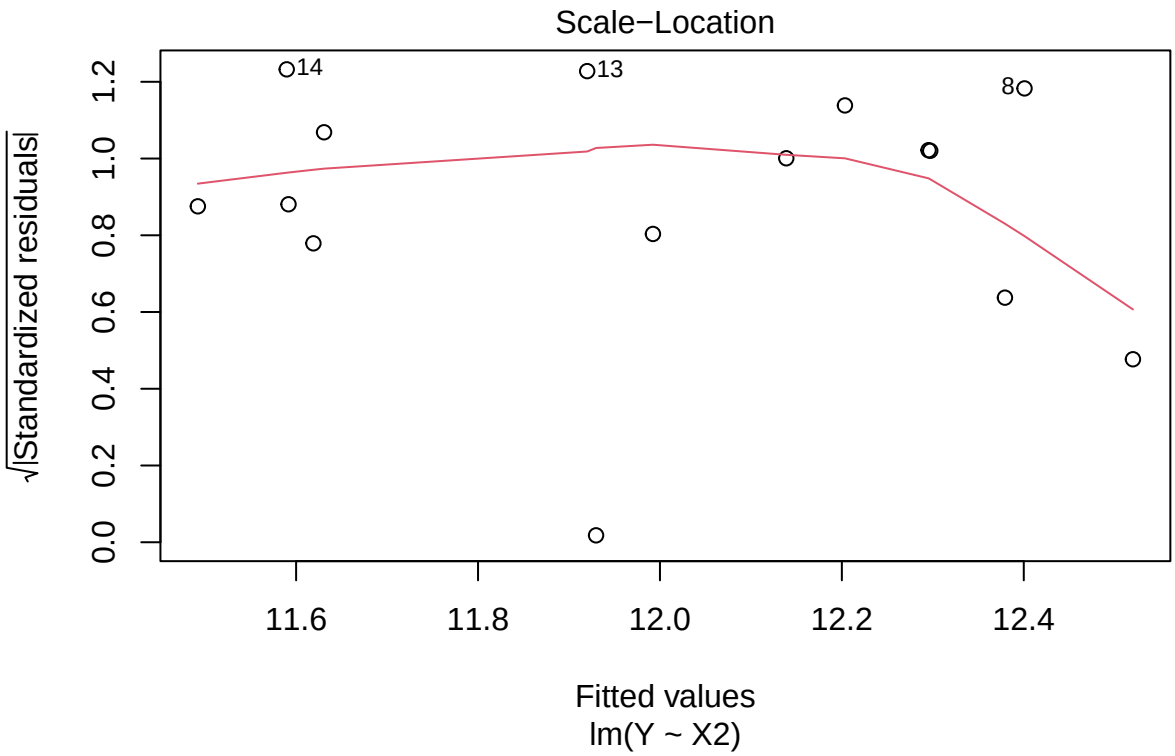

Scale−Location
√|Standardized residuals|
0.0
0.2
0.4
0.6
0.8
1.0
1.2
13
7
10
11.997
11.998
11.999
12.000
12.001
12.002
12.003
Fitted values
lm(Y ~ X1)

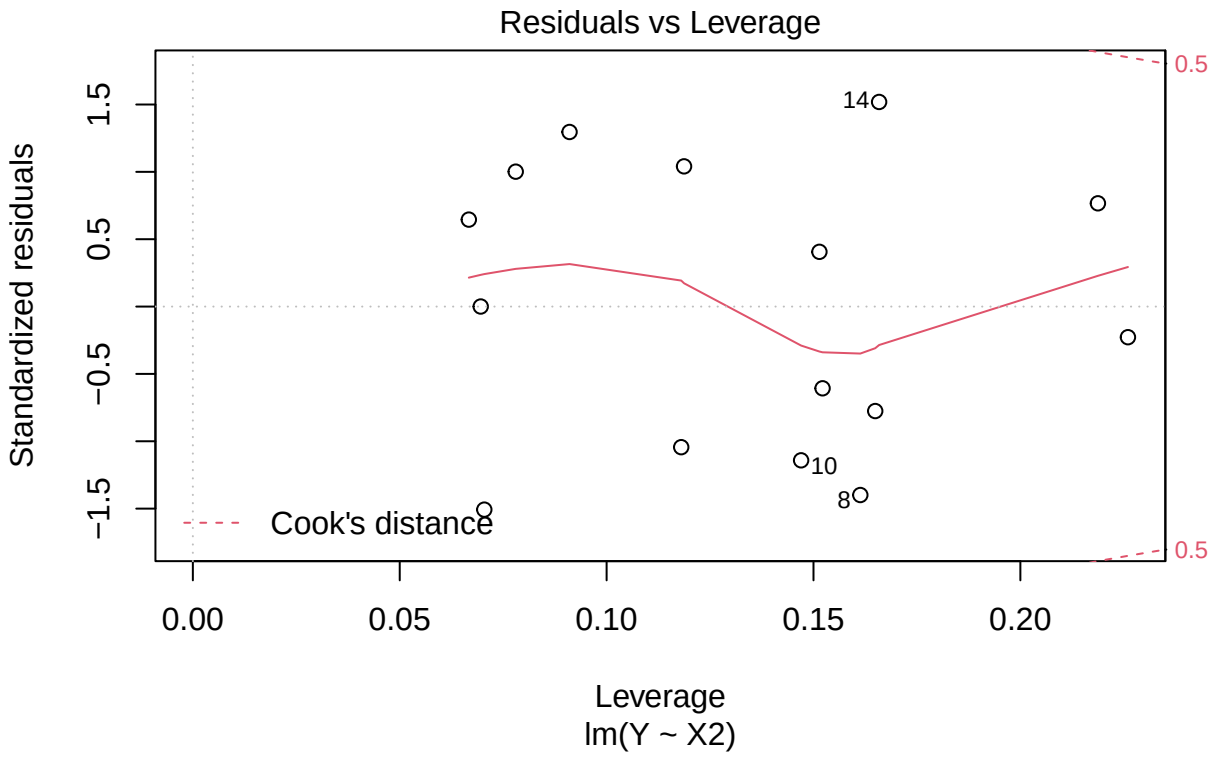

Residuals vs Leverage
Standardized residuals
1.5
0.5
−0.5
−1.5
0.5
0.5
14
8
10
Cook's distance
0.00
0.05
0.10
0.15
0.20
0.25
Leverage
lm(Y ~ X1)

```
plot(model2)
```

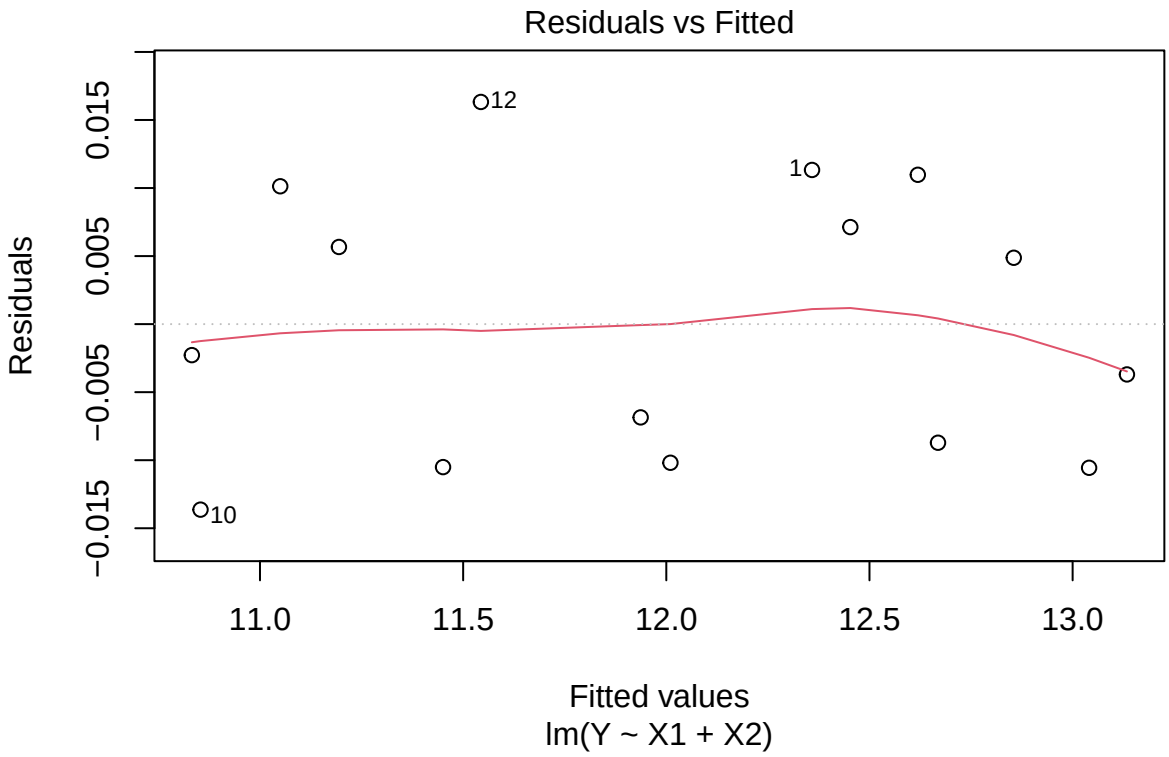

Residuals vs Fitted
Residuals
1.0
0.5
0.0
−0.5
−1.0
14
13
8
11.6
11.8
12.0
12.2
12.4
Fitted values
lm(Y ~ X2)

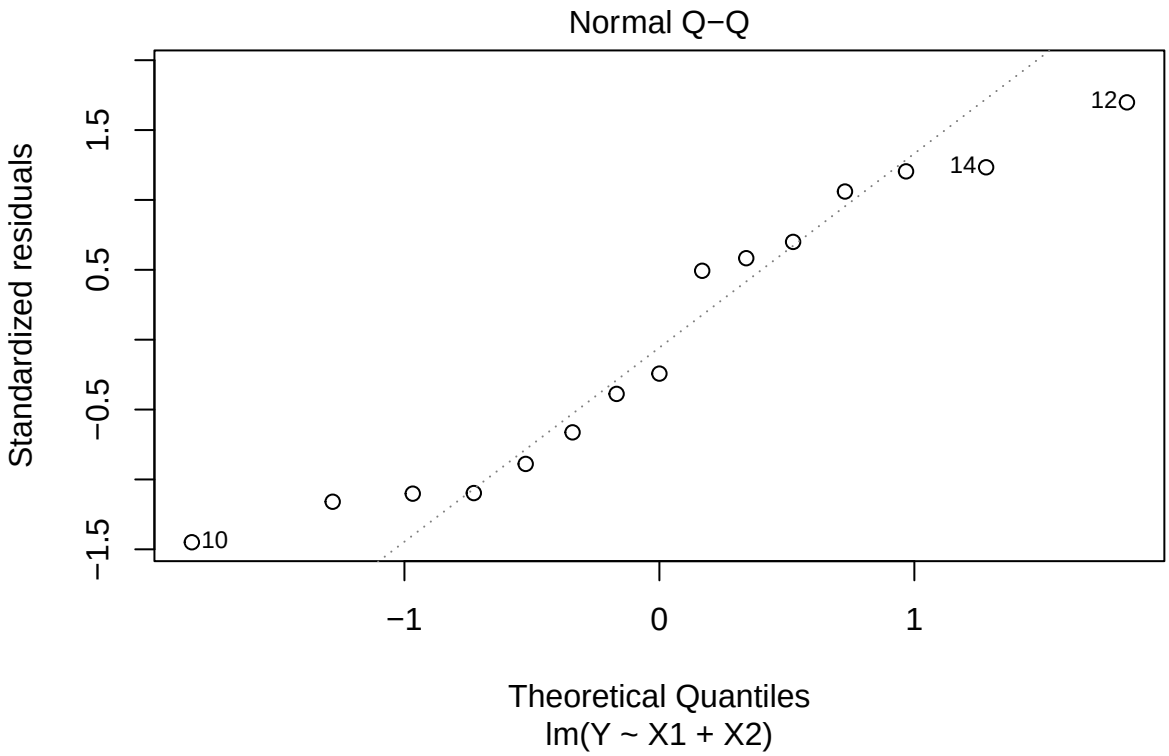

Normal Q−Q
Standardized residuals
−1.5
−0.5
0.5
1.0
1.5
13
8
14
−1
0
1
Theoretical Quantiles
lm(Y ~ X2)

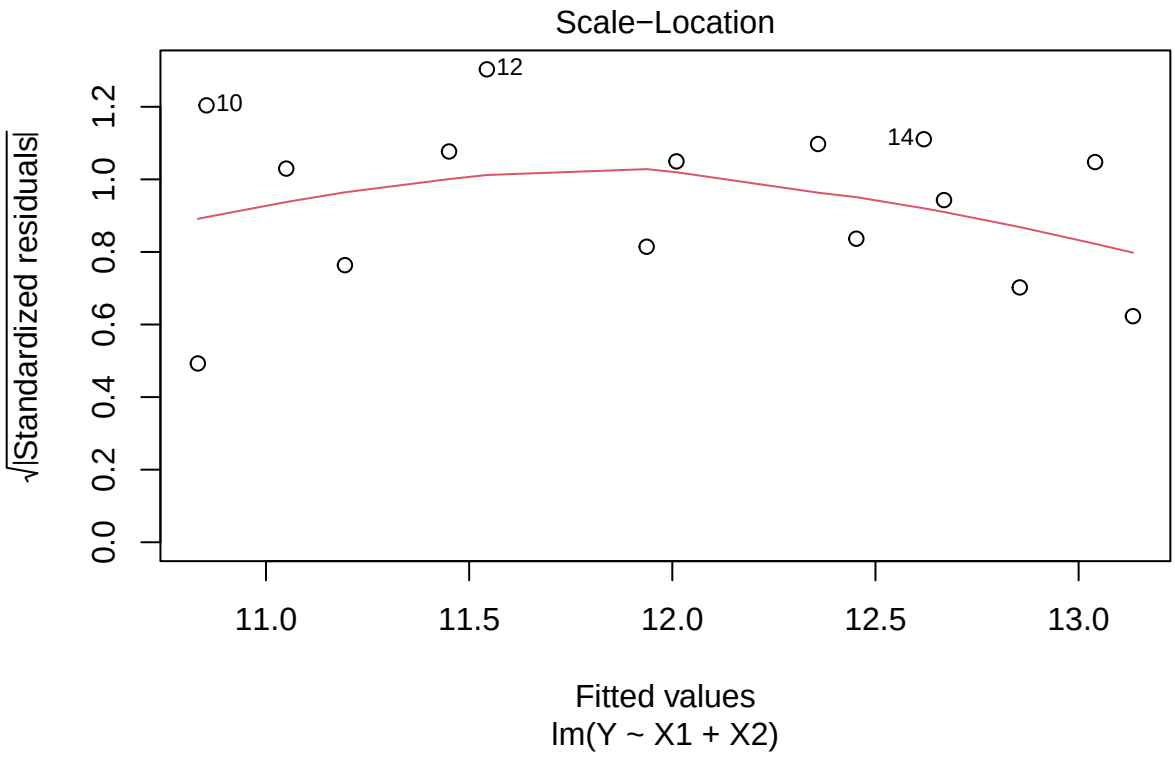

Scale−Location
√|Standardized residuals|
0.0
0.2
0.4
0.6
0.8
1.0
1.2
14
13
8
11.6
11.8
12.0
12.2
12.4
Fitted values
lm(Y ~ X2)

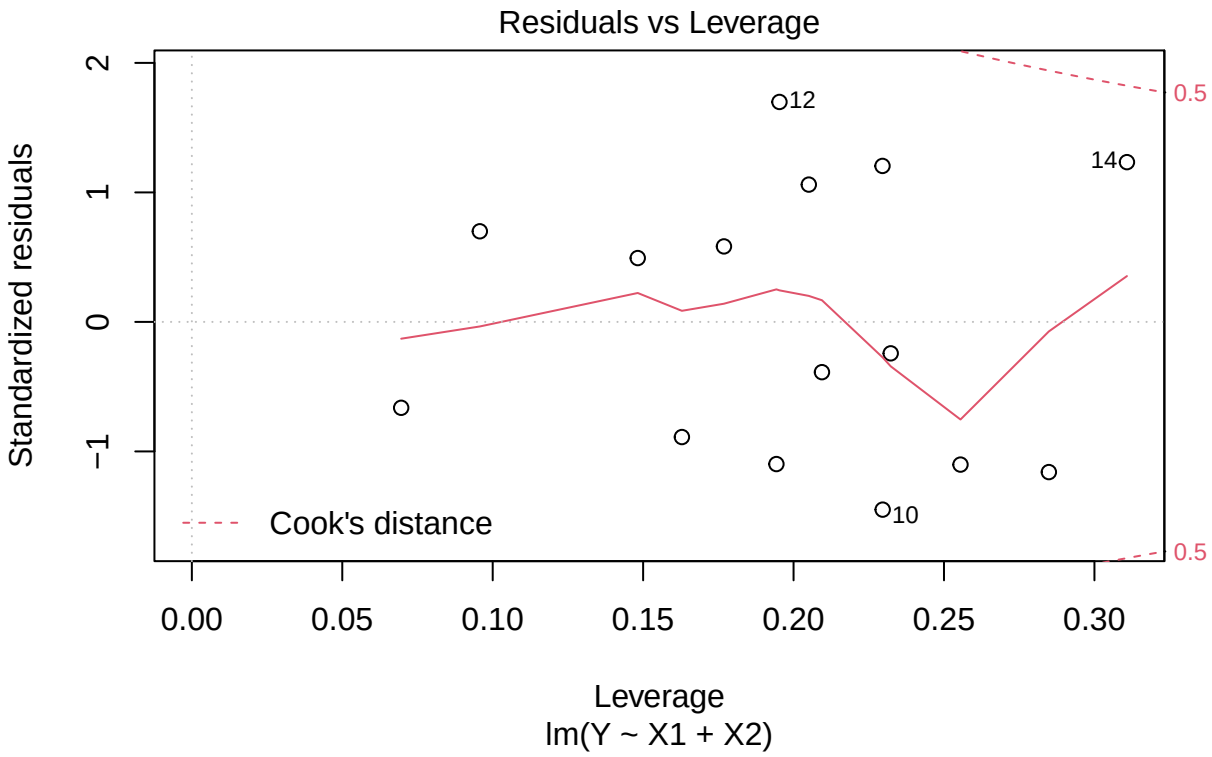



```
plot(model12)
```


Residuals vs Fitted
Residuals
0.015
0.005
−0.005
−0.015
12
1
10
11.0
11.5
12.0
12.5
13.0
Fitted values
lm(Y ~ X1 + X2)


Normal Q−Q
Standardized residuals
1.5
0.5
−0.5
−1.5
10
12
14
−1
0
1
Theoretical Quantiles
lm(Y ~ X1 + X2)


Scale−Location
√|Standardized residuals|
0.0
0.2
0.4
0.6
0.8
1.0
1.2
10
12
14
11.0
11.5
12.0
12.5
13.0
Fitted values
lm(Y ~ X1 + X2)



a) Report and Summaries provided in the above code.

b) The simple linear regression models are similar. They have similar, small slopes and intercepts near 10. The multiple regression model has coefficients in the whole numbers with a negative intercept. This is quite different from the simple linear regression models and is unusual. Additionally, the p-value for the X1 variable was a large .993 meaning that the variable is not statistically significant. The p-value for the X2 variable was also not statistically significant, but was closer to significance at a value of .106. The unusual aspect is that when both variables are used in the multiple linear regression model, their p-values reduce to <2e-16, making them both statistically significant. This is unusual because alone, the variables do not produce a productive model, but together they create a robust, statistically significant model.

c) I created residual plots in efforts to determine if any of the plots fail any conditions. I determined a Residuals vs Fitted graph showed the nature of the residual plots the best and they show there is a lack of linearity in the plots for the X1 and X2 variables. However, when the Residuals vs Fitted graph is used for the multiple linear regression model, the plot has a nice linear nature. All three of the plots pass the normality assumption because their Q-Q plots appear to be straight. Looking at the Scale-Location plots, some degree of fanning out/thickening is visible for the X1 and X2 variables. Considering this, the equal variance condition is not passed. However, if we look at the Scale-Location plot for the multiple linear regression model, the data plot appears to be straight and thus passes the equal variance condition. Using these plots, we can further our belief that individually, the X1 and X2 variables do not produce a statisfically significant model. When used together in a multiple linear regression model, the model is productive and and robust.