C1M5_peer_reviewed

January 13, 2022

1 Module 5: Peer Reviewed Assignment

1.0.1 Outline:

The objectives for this assignment:

1. Understand what can cause violations in the linear regression assumptions.
2. Enhance your skills in identifying and diagnosing violated assumptions.
3. Learn some basic methods of addressing violated assumptions.

General tips:

1. Read the questions carefully to understand what is being asked.
2. This work will be reviewed by another human, so make sure that you are clear and concise in what your explanations and answers.

[38]:

```
# Load Required Packages
library(ggplot2)
```

1.1 Problem 1: Let's Violate Some Assumptions!

When looking at a single plot, it can be difficult to discern the different assumptions being violated. In the following problem, you will simulate data that purposefully violates each of the four linear regression assumptions. Then we can observe the different diagnostic plots for each of those assumptions.

1. (a) Linearity Generate SLR data that violates the linearity assumption, but maintains the other assumptions. Create a scatterplot for these data using ggplot.

Then fit a linear model to these data and comment on where you can diagnose nonlinearity in the diagnostic plots.

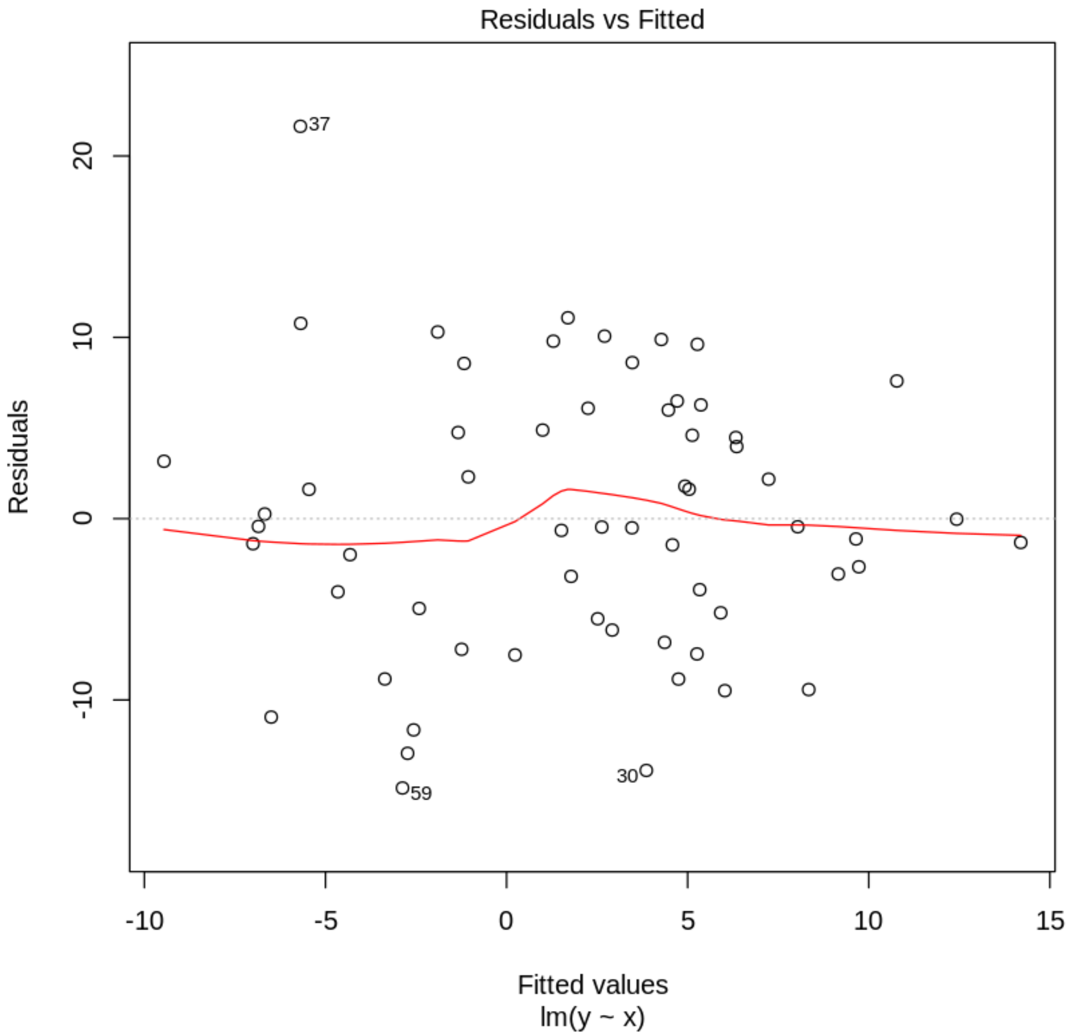
[39]:

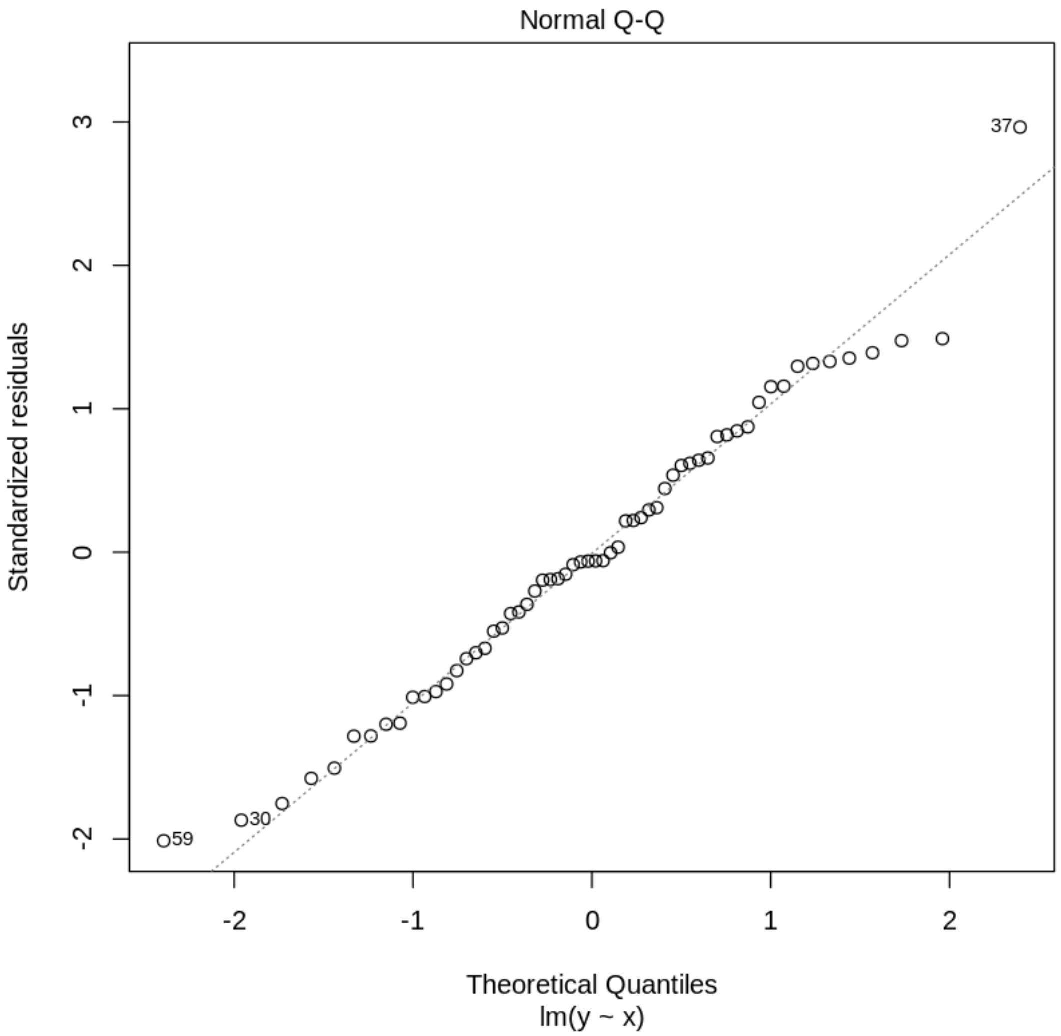
```
# Your Code Here
mu.1a = 0
var.1a = 2
x.1a = rnorm(60, mu.1a, var.1a)
err = rnorm(60, 0, 8)
y.1a = (2 * x.1a) / 3 + x.1a + err
```

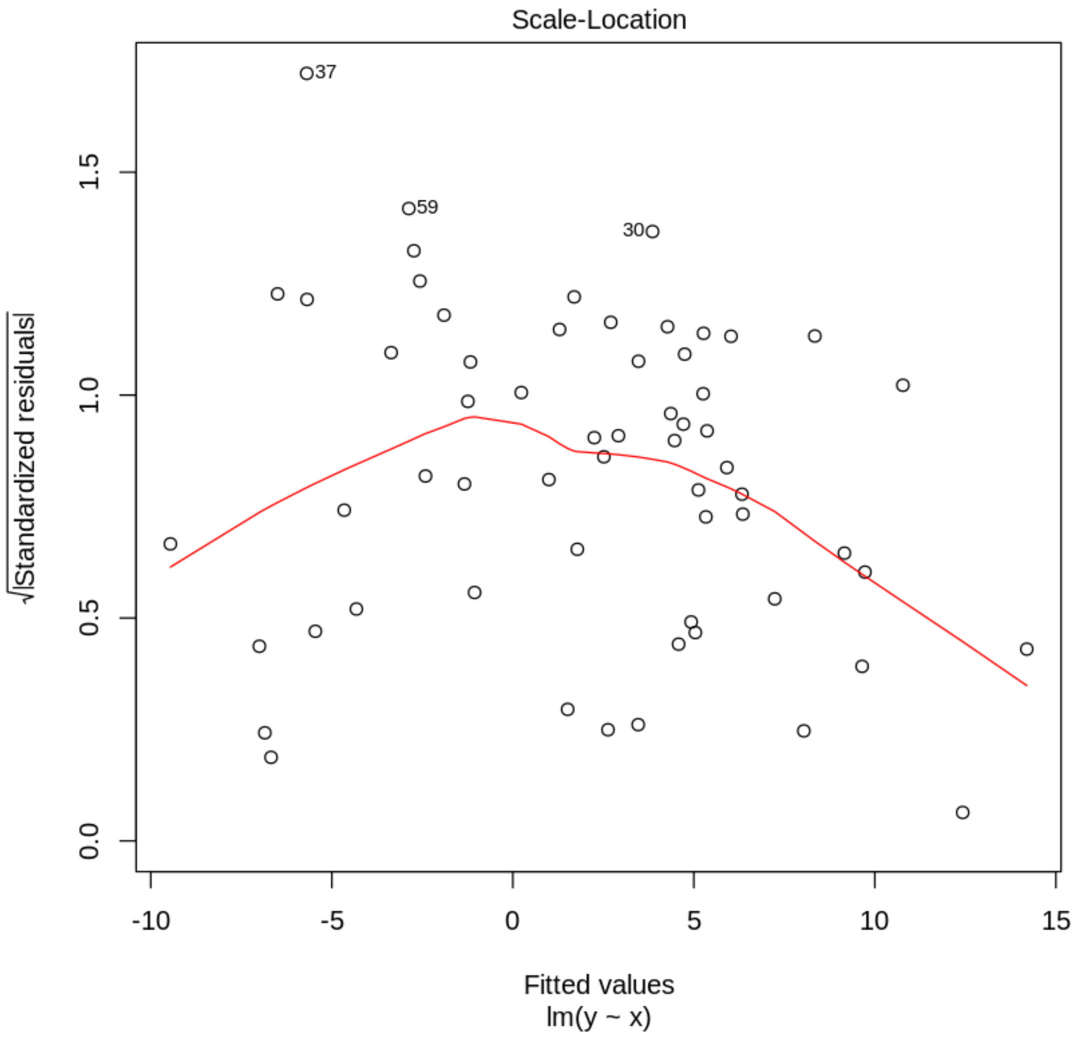
```
df.1a = data.frame(x=x.1a, y=y.1a)

lm.1a = lm(y ~ x, data=df.1a)

df.diagnostics.1a = data.frame(yhat = fitted(lm.1a), res = resid(lm.1a), y=df.
 ↪1a$y)
```

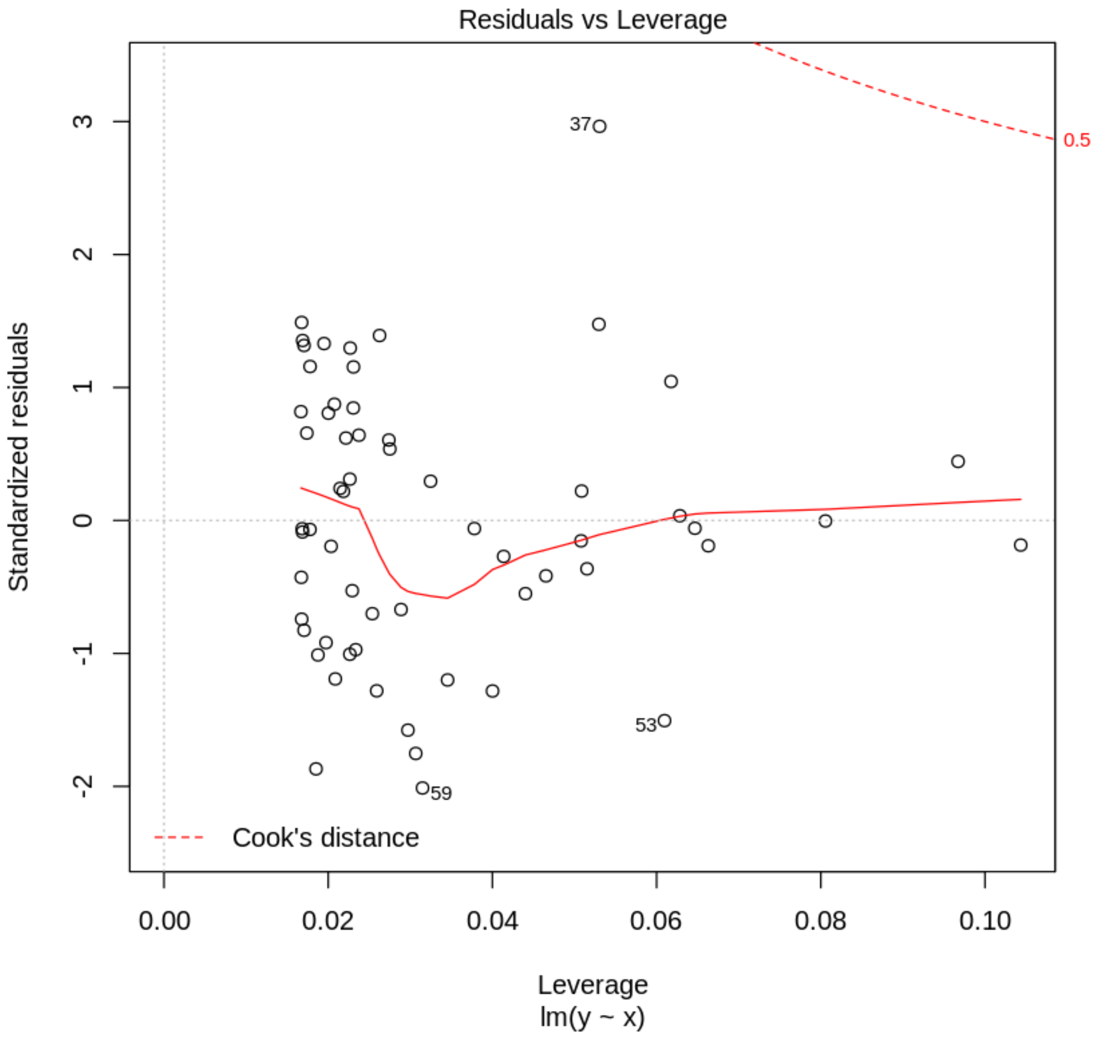
[40]:

```
plot(lm.1a)
```




Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
37
30
59
-2
-1
0
1
2
Theoretical Quantiles
lm(y ~ x)


Scale-Location
37
59
30
√|Standardized residuals|
0.0
0.5
1.0
1.5
-10
-5
0
5
10
15
Fitted values
lm(y ~ x)



This data violates the linearity assumption. In this oberved values vs fitted plot doesn't follow the 'y=x' line on the qq plot. However, we can see that we have constant variance by the fitted values vs residuals plots. Also, our simulated data shows error independence and normally distributed.

1. (b) Homoskedasticity Simulate another SLR dataset that violates the constant variance assumption, but maintains the other assumptions. Then fit a linear model to these data and comment on where you can diagnose non-constant variance in the diagnostic plots.

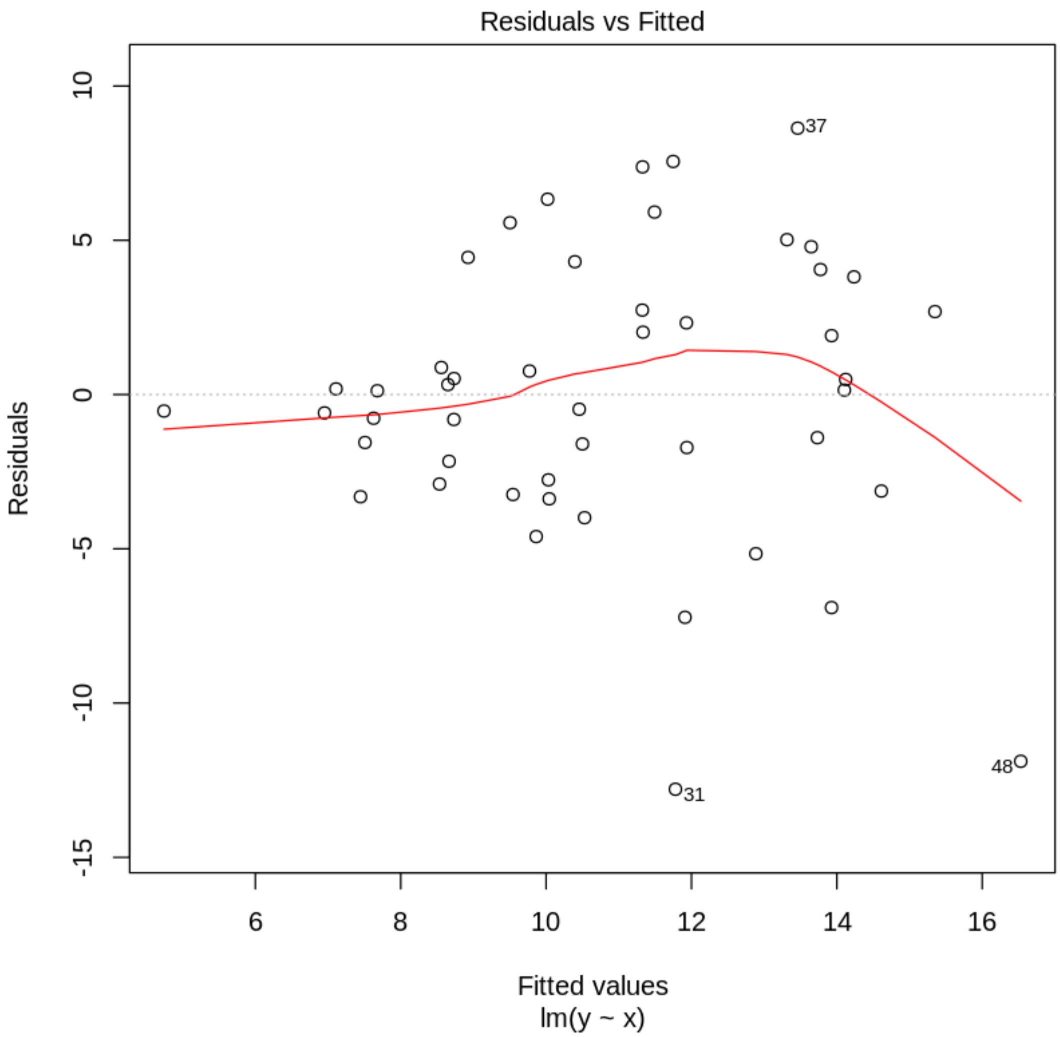
[41]:

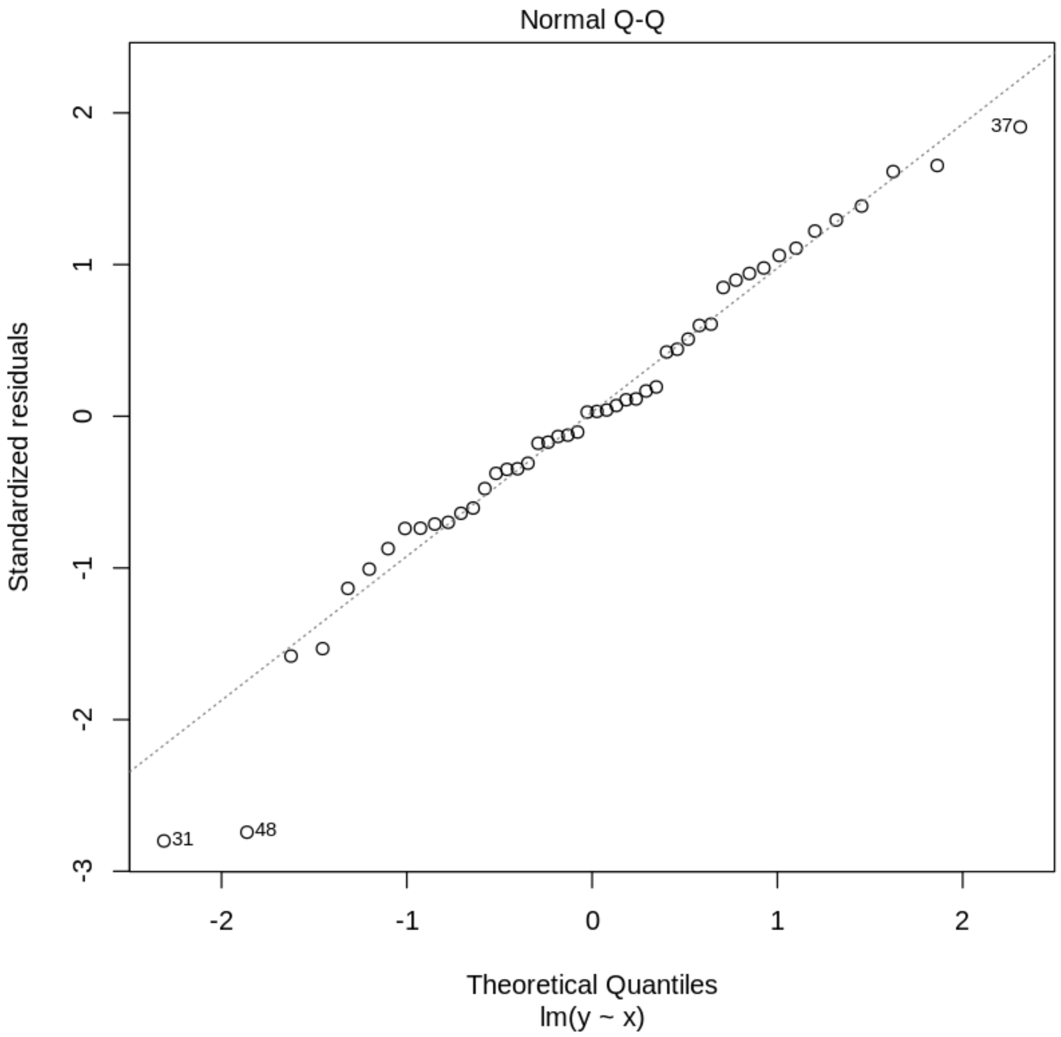
```
# Your Code Here
x.1b = rnorm(48, 0, 3)
x.1b = sort(x.1b)
err = c()
err[1:12] = rnorm(12, sd=1)
```

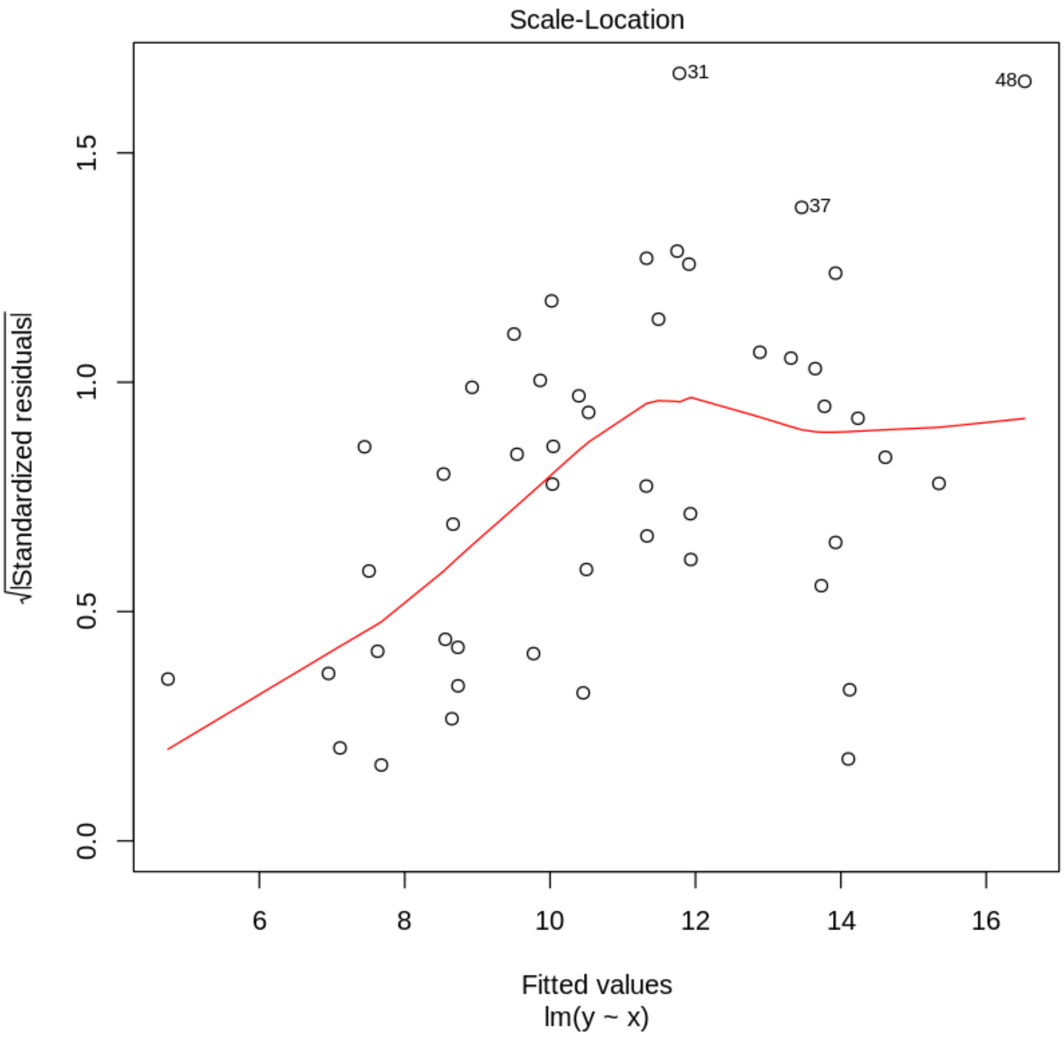
```
err[13:24] = rnorm(12, sd=3)
err[25:36] = rnorm(12, sd=6)
err[37:48] = rnorm(12, sd=9)
y.1b = 10 + x.1b + err

df.1b = data.frame(x=x.1b, y=y.1b)

lm.1b = lm(y ~ x, data=df.1b)

df.diagnostics.1b = data.frame(yhat = fitted(lm.1b), res = resid(lm.1b), y=df.
 ↪1b$y)
```

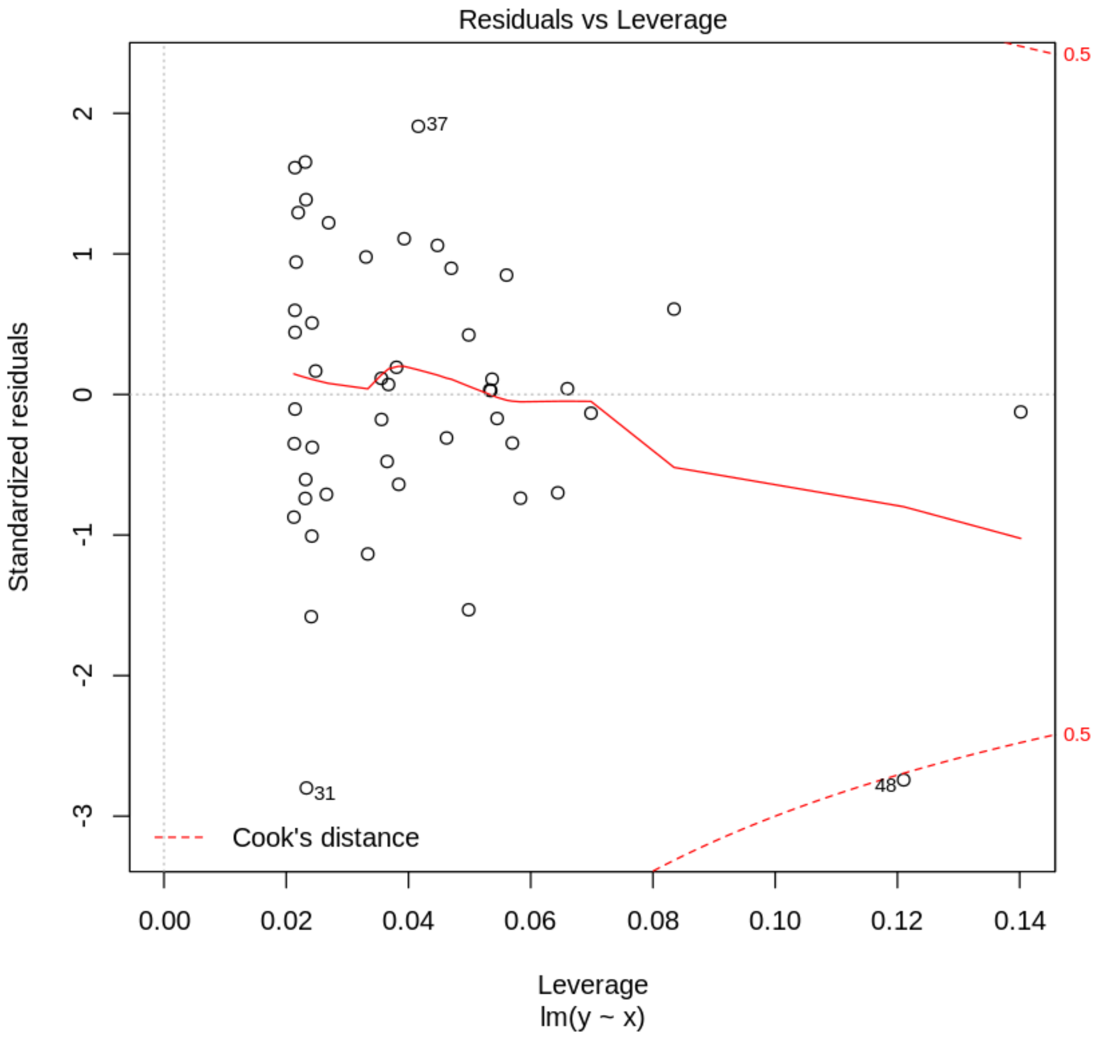
[42]:

```
plot(lm.1b)
```



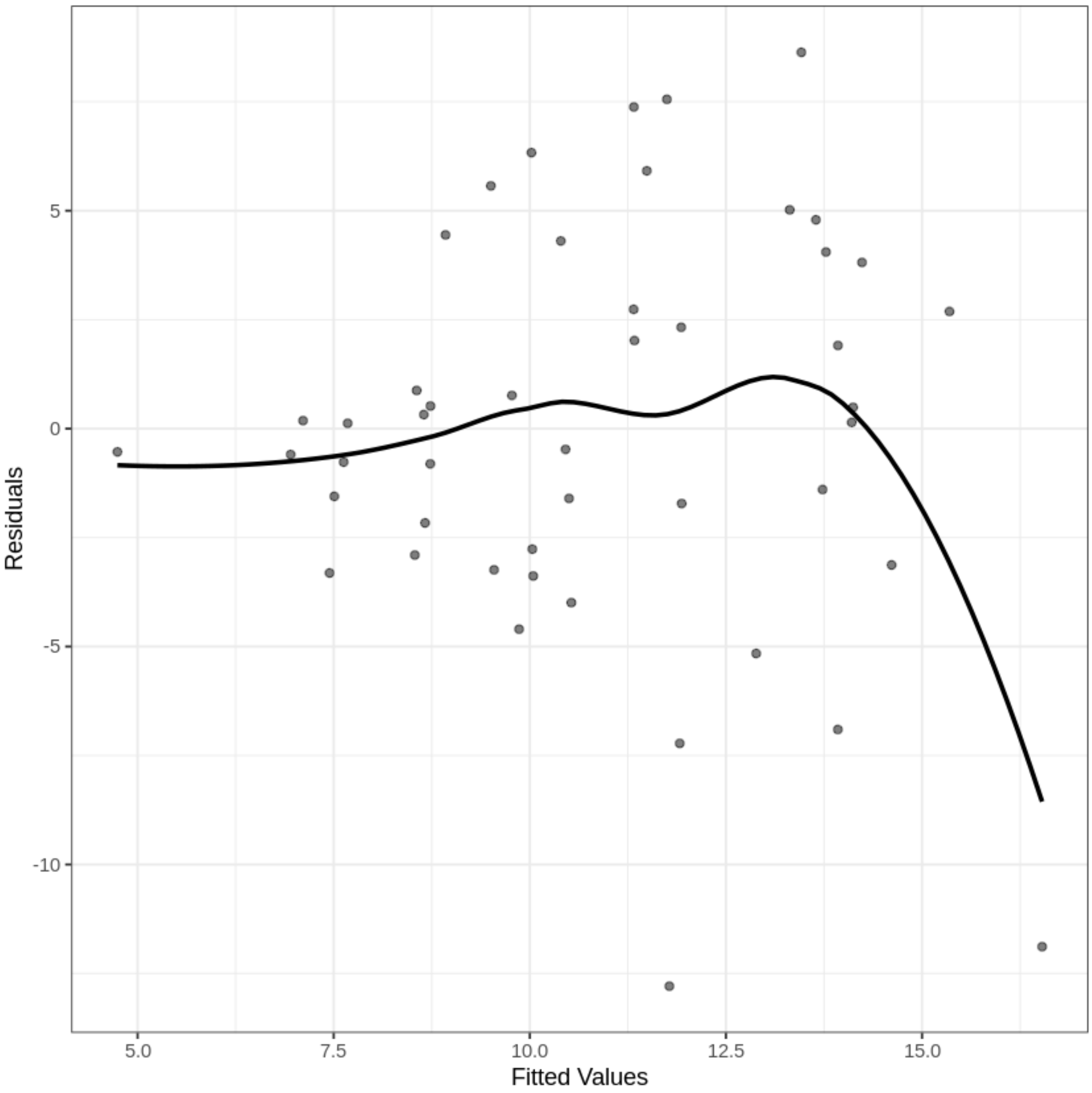

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(y ~ x)
-3
-2
-1
0
1
2
-2
-1
0
1
2
31
48
37


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
6
8
10
12
14
16
Fitted values
lm(y ~ x)
31
48
37



```
[43]: ggplot(df.diagnostics.1b, aes(x = yhat, y = res)) +
       geom_point(alpha = 0.5) +
       geom_smooth(se = F, col = "black") +
       xlab("Fitted Values") + ylab("Residuals")+
       theme_bw()
```

```
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



We can see the residulas become larger as the fited values become larger. Our linearity remains the same as shown by the qqplot, the errors are independent and are normally distributed as shown by our simulation.

1. (c) Independent Errors Repeat the above process with simulated data that violates the independent errors assumption.

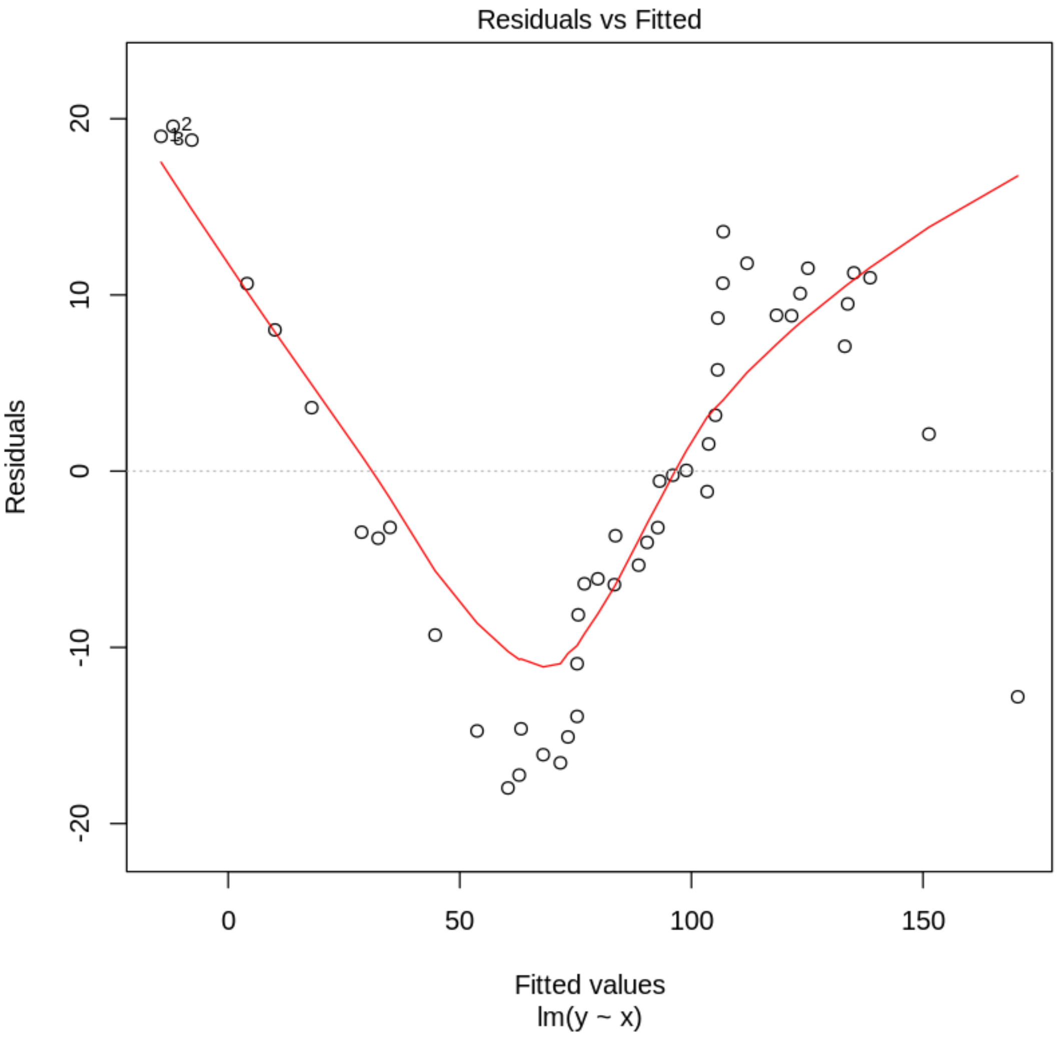
[44]:

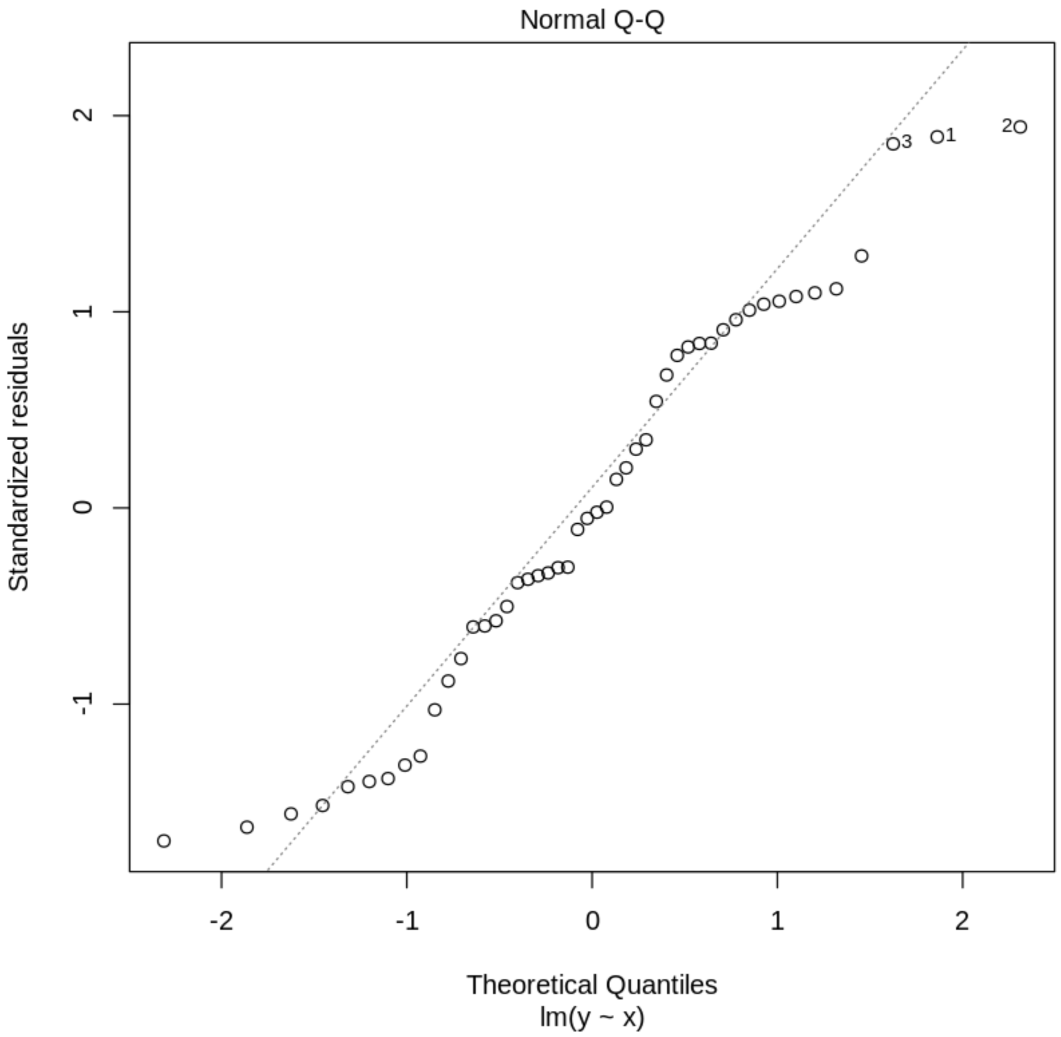
```
# Your Code Here
# Your Code Here
x.1c = rnorm(48, 0, 3)
x.1c = sort(x.1c)
err = rep(0,48)
err[1] = 1
```

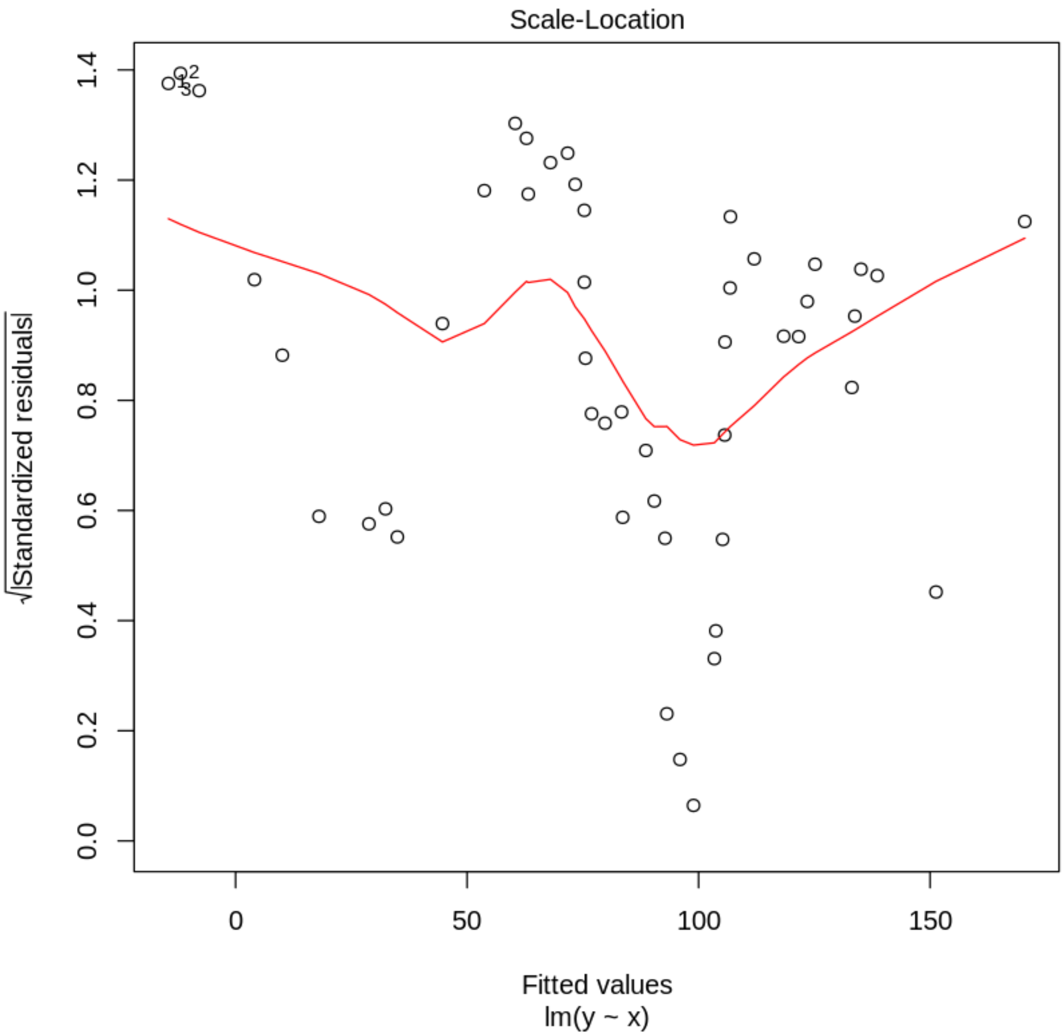
```
i=1
for (error in err){
    if(i==1){
        i = i +1
        next
    }
    err[i] = err[i -1] + 3
    i = i + 1
}

y.1c = 10 + x.1c + err

df.1c = data.frame(x=x.1c, y=y.1c)

lm.1c = lm(y ~ x, data=df.1c)

df.diagnostics.1c = data.frame(yhat = fitted(lm.1c), res = resid(lm.1c), y=df.
 ↪1c$y)
```

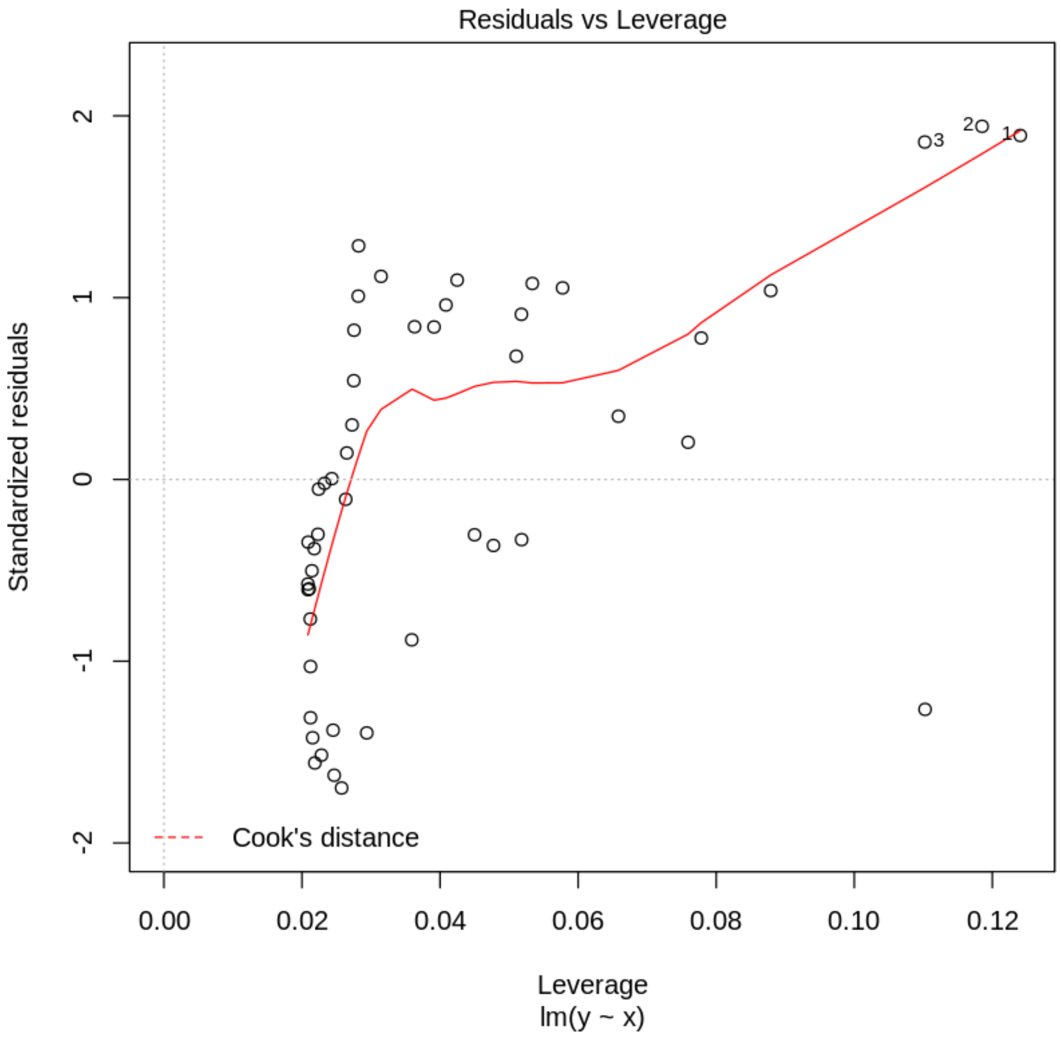
[45]:

```
plot(lm.1c)
```


Residuals vs Fitted
Residuals
20
10
0
-10
-20
0
50
100
150
Fitted values
lm(y ~ x)
1
2
3


Normal Q-Q
Standardized residuals
2
1
0
-1
-2
-1
0
1
2
Theoretical Quantiles
lm(y ~ x)
3
1
2

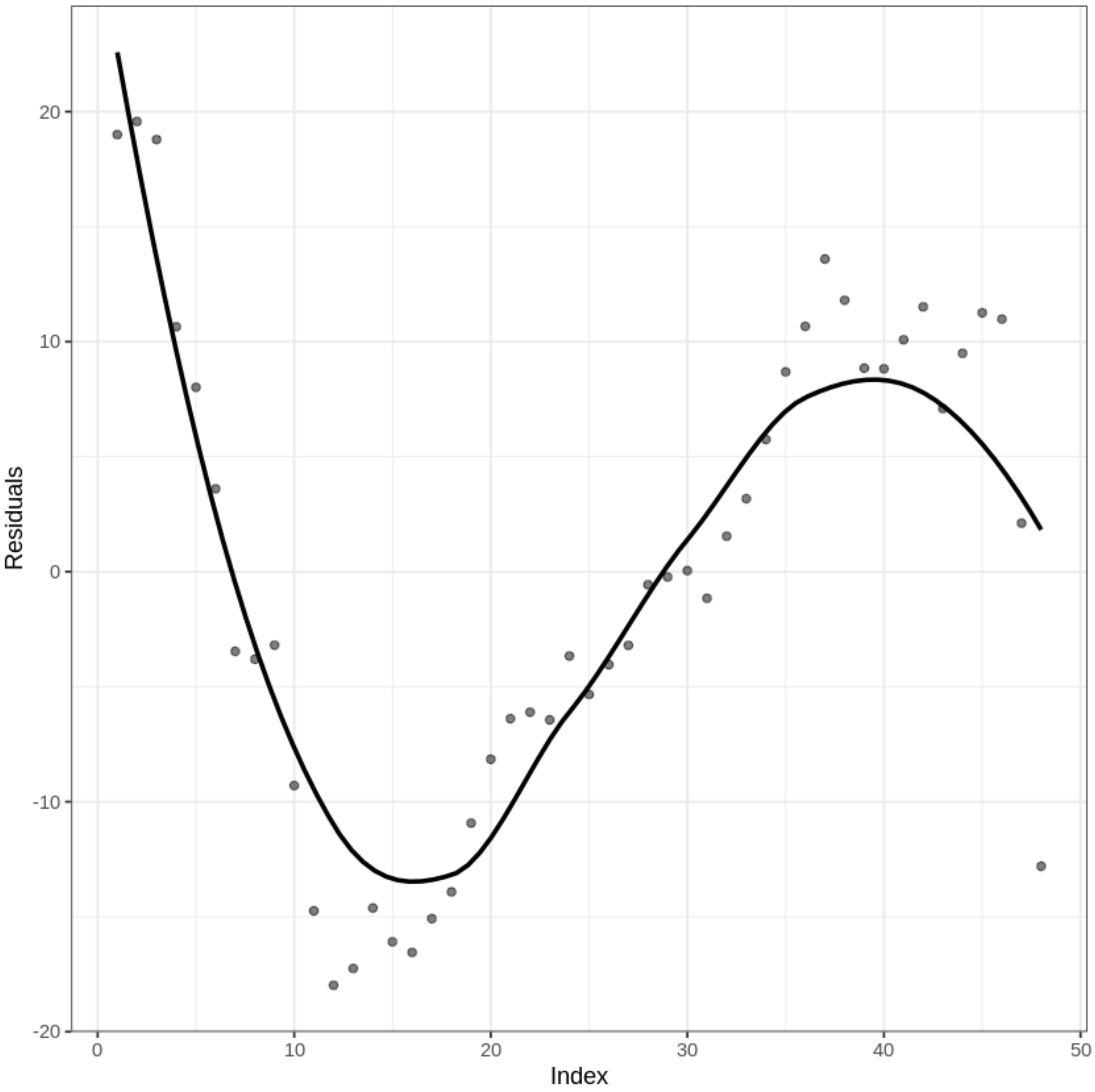

Scale-Location
√|Standardized residuals|
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
50
100
150
Fitted values
lm(y ~ x)
1
2
3



[46]:

```
ggplot(df.diagnostics.1c, aes(x = 1:length(df.diagnostics.1c$y), y = res)) +
  geom_point(alpha = 0.5) +
  xlab("Index") +
  geom_smooth(se = F, col = "black") +
  ylab("Residuals") +
  theme_bw()
```

```
Warning message:
"Use of `df.diagnostics.1c$y` is discouraged. Use `y` instead."
Warning message:
"Use of `df.diagnostics.1c$y` is discouraged. Use `y` instead."
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



If we had independence of errors, we would have randomly scattered errors around zero on this last plot. Given our simulation and qq plot, we know our data is linear; however, the dependence of errors is casuing issues.

1. (d) Normally Distributed Errors Only one more to go! Repeat the process again but simulate the data with non-normal errors.

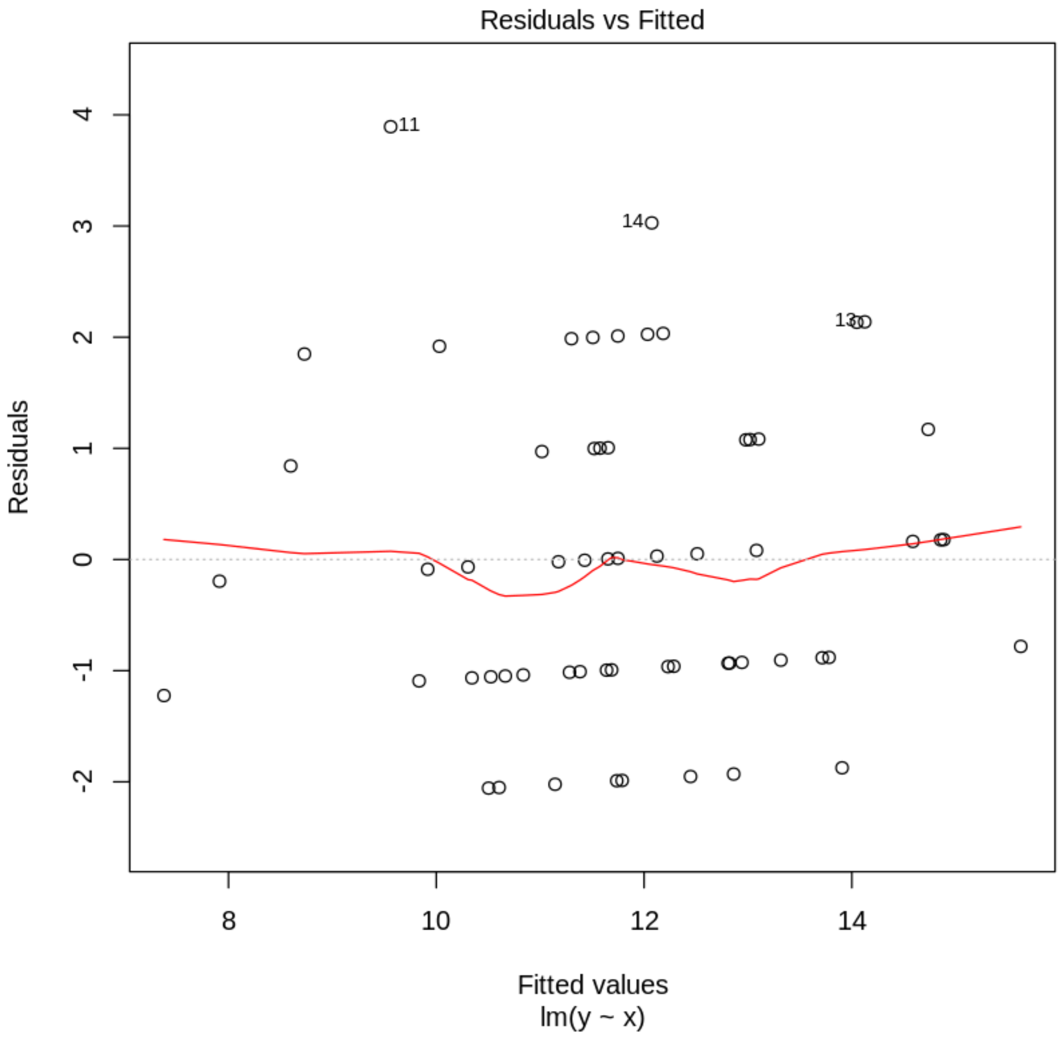
[47]:

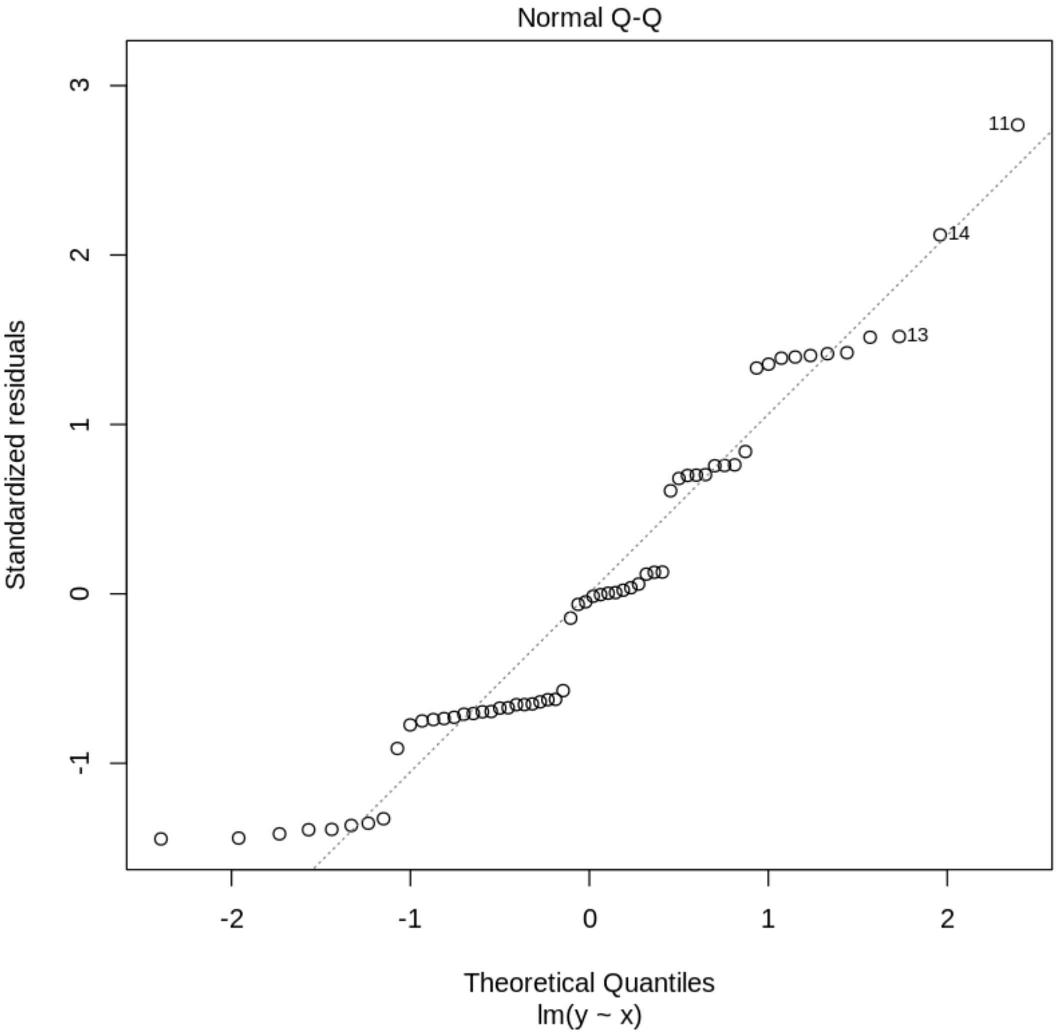
```
# Your Code Here
# Your Code Here
mu.1d = 0
var.1d = 2
x.1d = rnorm(60, mu.1d, var.1d)
err.1d = rpois(60, 2)
```

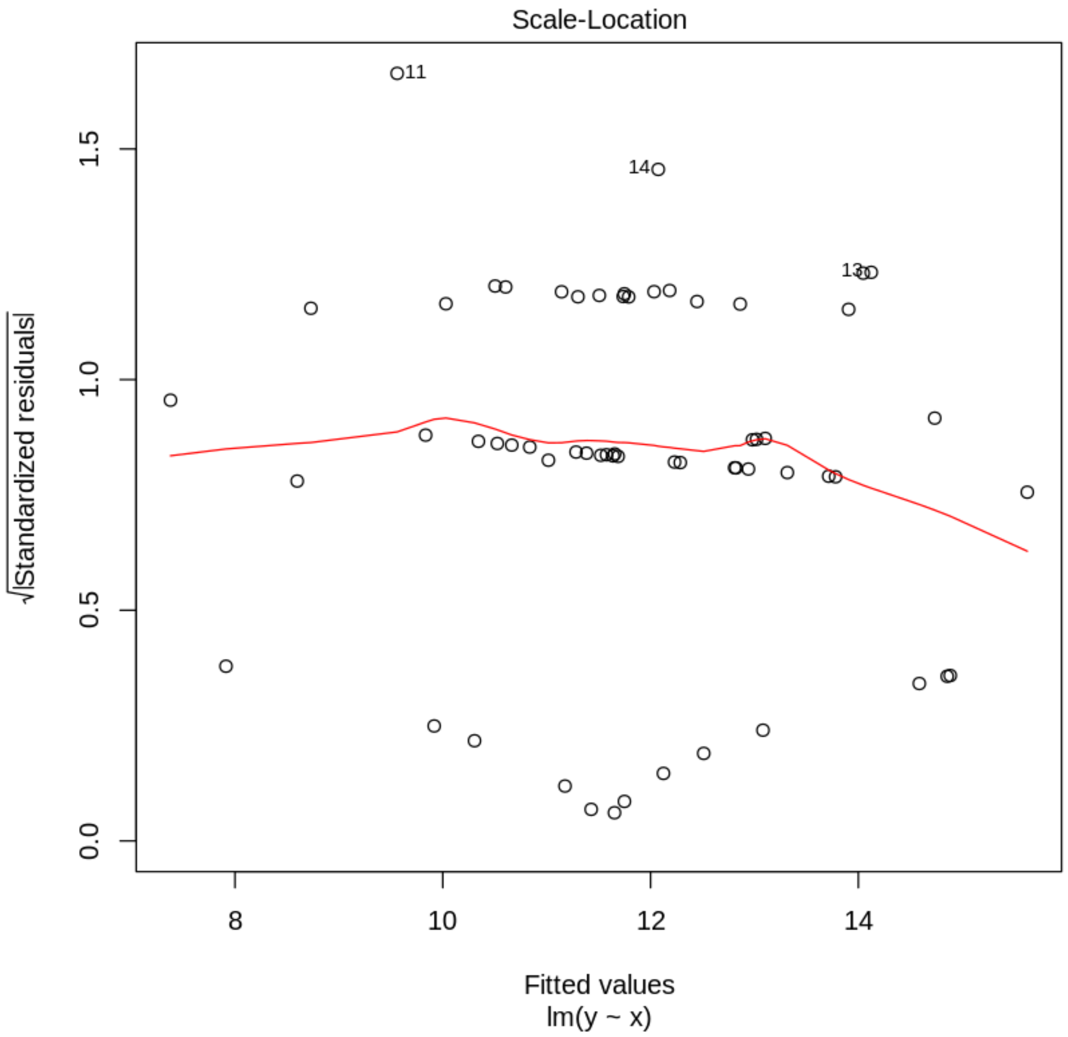
```
y.1d = 10 + x.1d + err.1d
df.1d = data.frame(x=x.1d, y=y.1d)

lm.1d = lm(y ~ x, data=df.1d)

df.diagnostics.1d = data.frame(yhat = fitted(lm.1d), res = resid(lm.1d), y=df.1d$y)
```

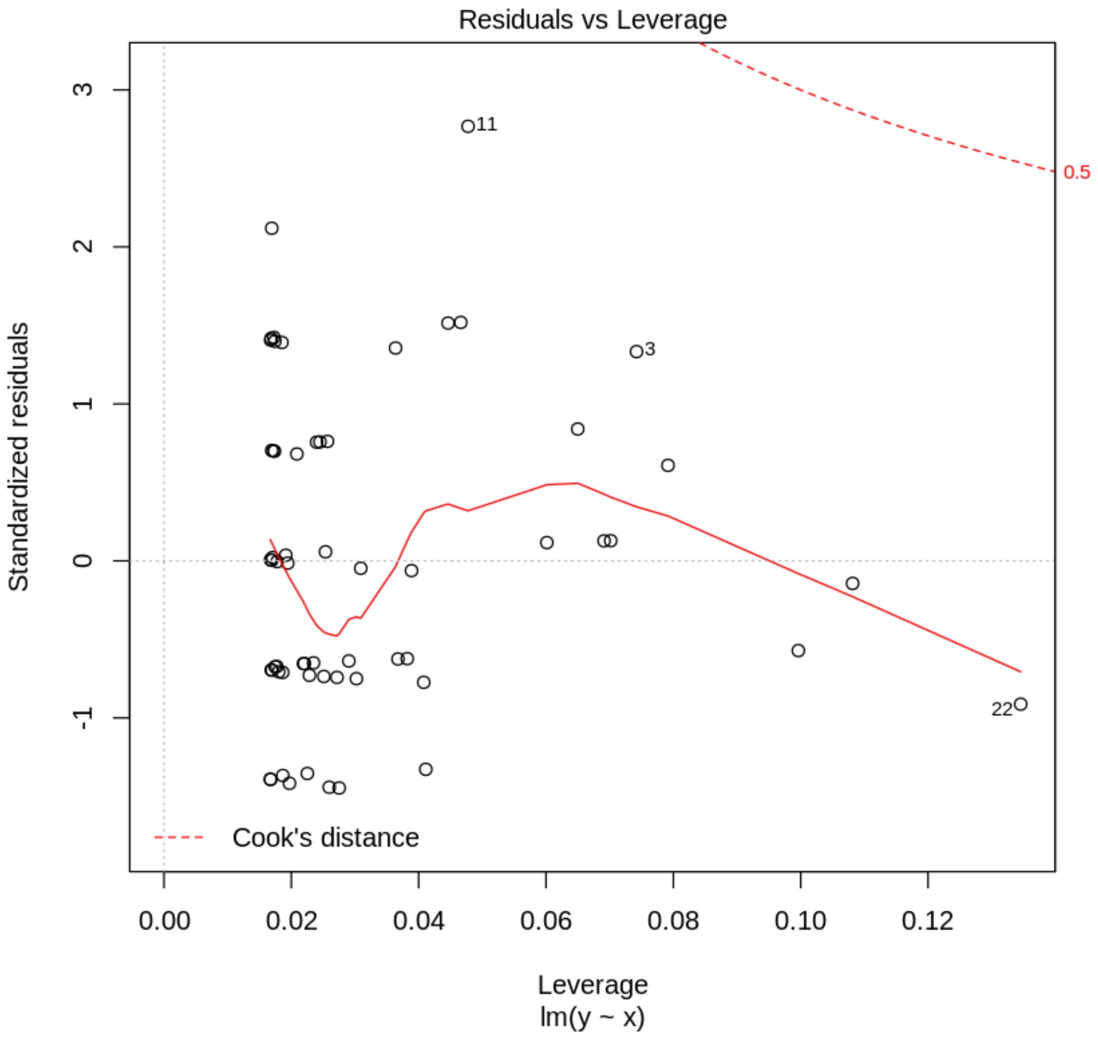
[48]:

```
plot(lm.1d)
```




Normal Q-Q
Standardized residuals
-1
0
1
2
3
-2
-1
0
1
2
Theoretical Quantiles
lm(y ~ x)
11
14
13


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
11
14
13
8
10
12
14
Fitted values
lm(y ~ x)



Using the QQ plot and the residuals vs fitted plots, we can tell see the normality issues whin the plot. The QQ plot doesn't follow the "y=x" line; however, no specific pattern is followed. Therefore, we can tell that there is a normality issues among the errors.

2 Problem 2: Hats for Sale

Recall that the *hat* or *projection* matrix is defined as

$$H = X(X^TX)^{-1}X^T.$$

The goal of this question is to use the hat matrix to prove that the fitted values, $\widehat{\mathbf{Y}}$, and the residuals, $\widehat{\varepsilon}$, are uncorrelated. It's a bit of a process, so we will do it in steps.

2. (a) Show that $\widehat{Y} = HY$. That is, H “puts a hat on” Y.

2. (b) Show that H is symmetric: $H = H^T$. For ease of typing, let ’ represent transpose. For example, H’ equals H^T.

$$H = X(X'X)^{-1}X' = X((X'X')')^{-1}X' = (X')'((X'X)^{-1})'X' = H'$$

2. (c) Show that $H(I_n - H) = 0_n$, where 0_n is the zero matrix of size $n \times n$.**

$$H(I_n - H) = H(-H) = 0_n$$

2. (d) Stating that $\widehat{\mathbf{Y}}$ is uncorrelated with $\widehat{\varepsilon}$ is equivalent to showing that these vectors are orthogonal.* That is, we want their dot product to equal zero:

$$\widehat{\mathbf{Y}}^T\widehat{\varepsilon} = 0.$$

Prove this result. Also explain why being uncorrelated, in this case, is equivalent to the being orthogonal.

2.(e) Why is this result important in the practical use of linear regression? This is the same as showing the reisduals have constannt variance and center around zero. If the dot product dooesn’t equal zero there is some issues with homeoskedaticity.

2.1 Problem 3: Model Diagnosis

We here at the University of Colorado’s Department of Applied Math love Bollywood movies. So, let’s analyze some data related to them!

We want to determine if there is a linear relation between the amount of money spent on a movie (it’s budget) and the amount of money the movie makes. Any venture capitalists among you will certianly hope that there is at least some relation. So let’s get to modelling!

3. (a) Initial Inspection Load in the data from local directory and create a linear model with `Gross` as the response and `Budget` as the feature. The data is stored in the same local directory and is called `bollywood_boxoffice.csv`. Thank the University of Florida for this specific dataset.

Specify whether each of the four regression model assumptions are being violated.

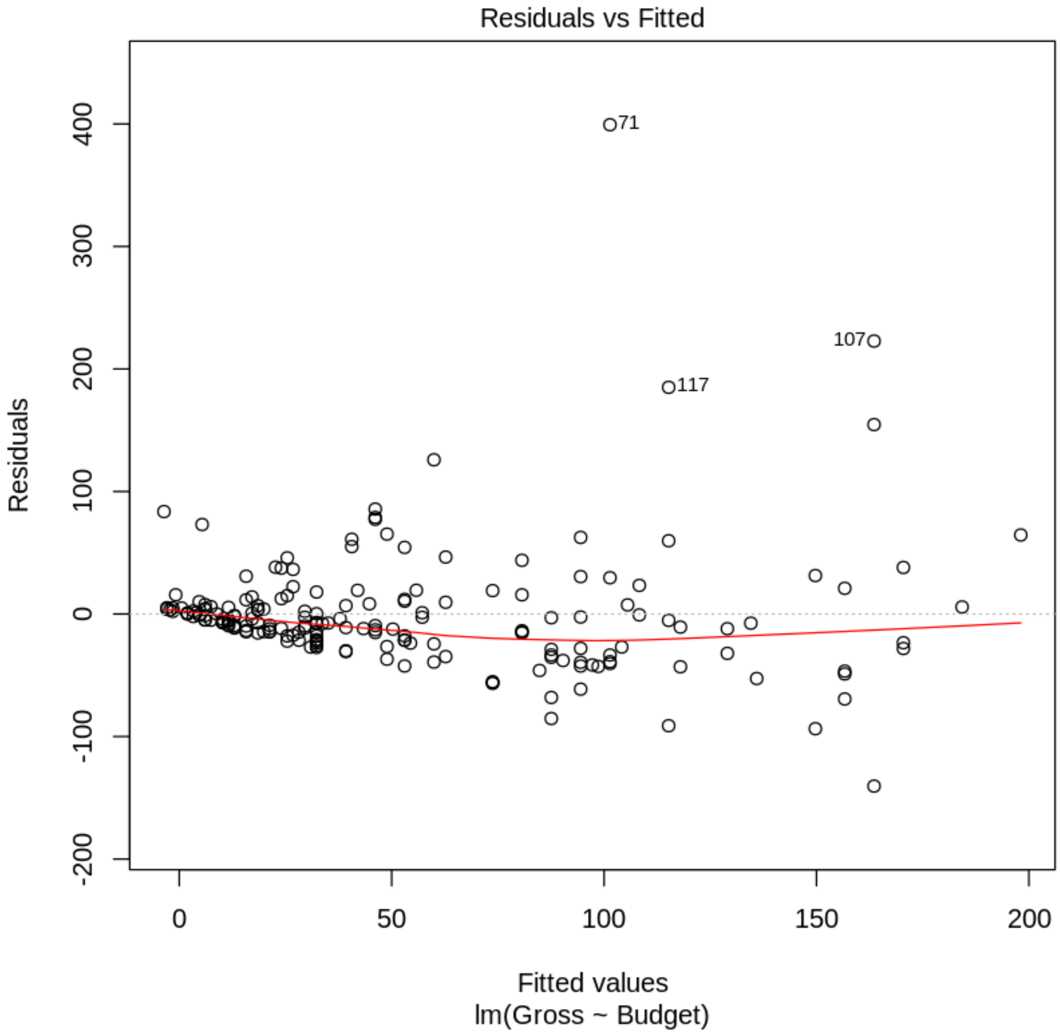
Data Source: http://www.bollymoviereviewz.com

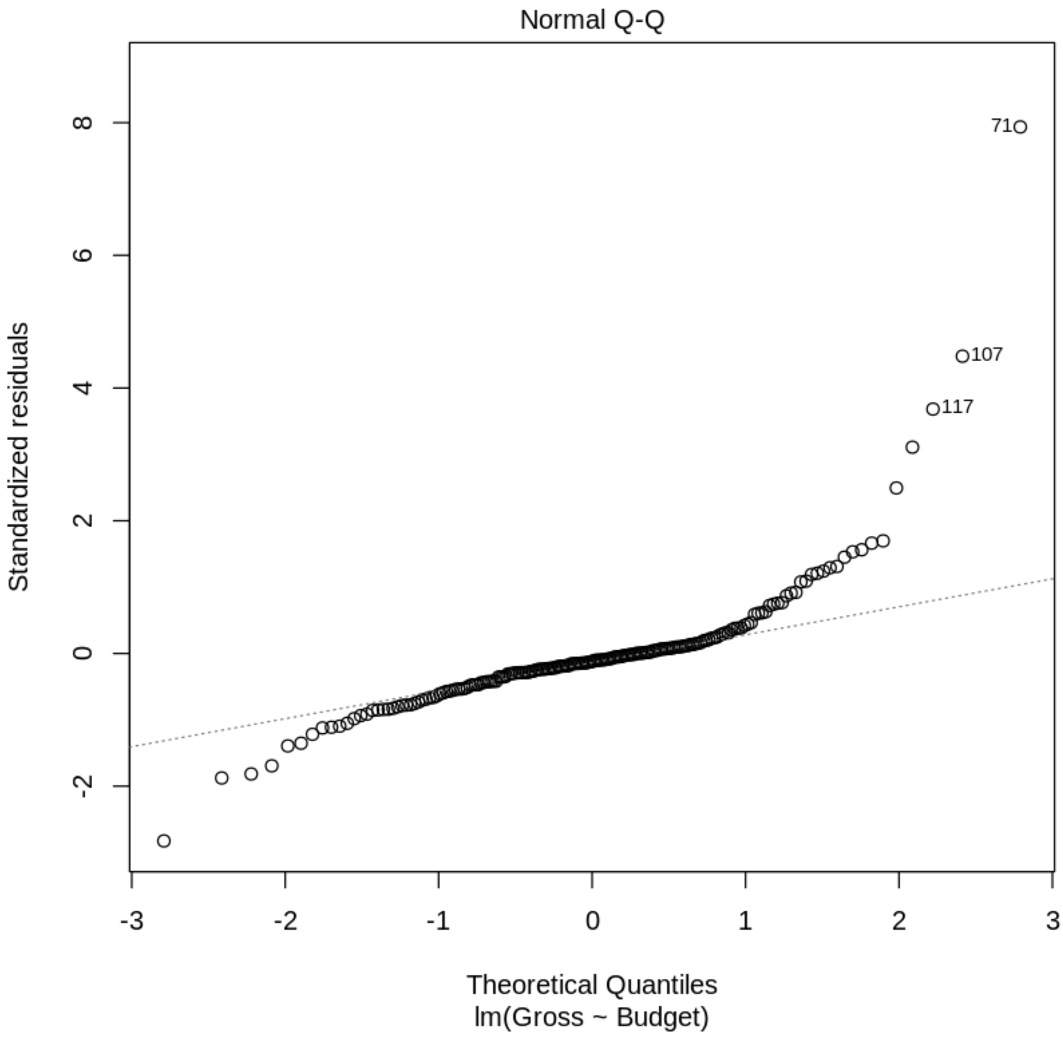
[49]:

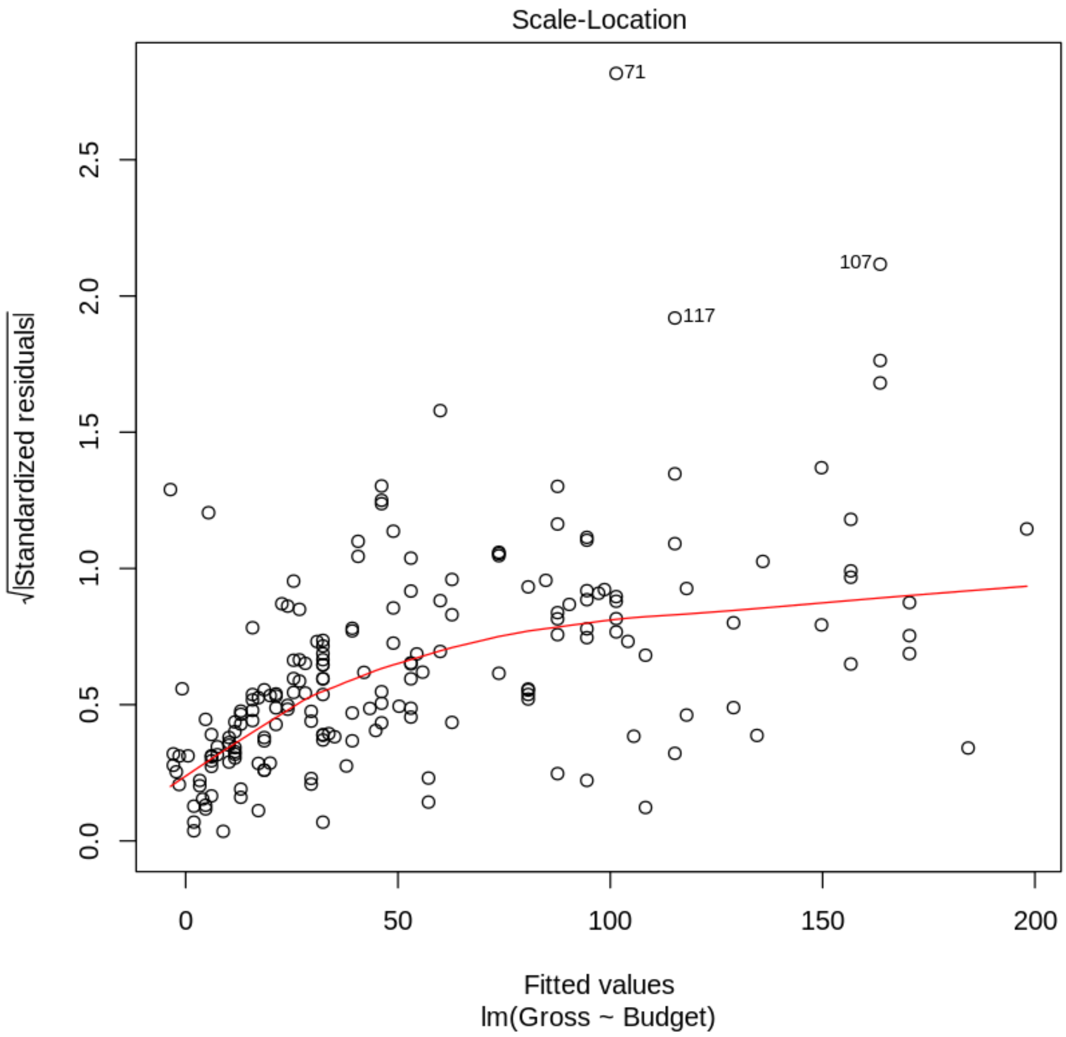
```
# Load the data
bollywood = read.csv("bollywood_boxoffice.csv")
summary(bollywood)

# Your Code Here
bolly.lm = lm(Gross~Budget, data=bollywood)
```

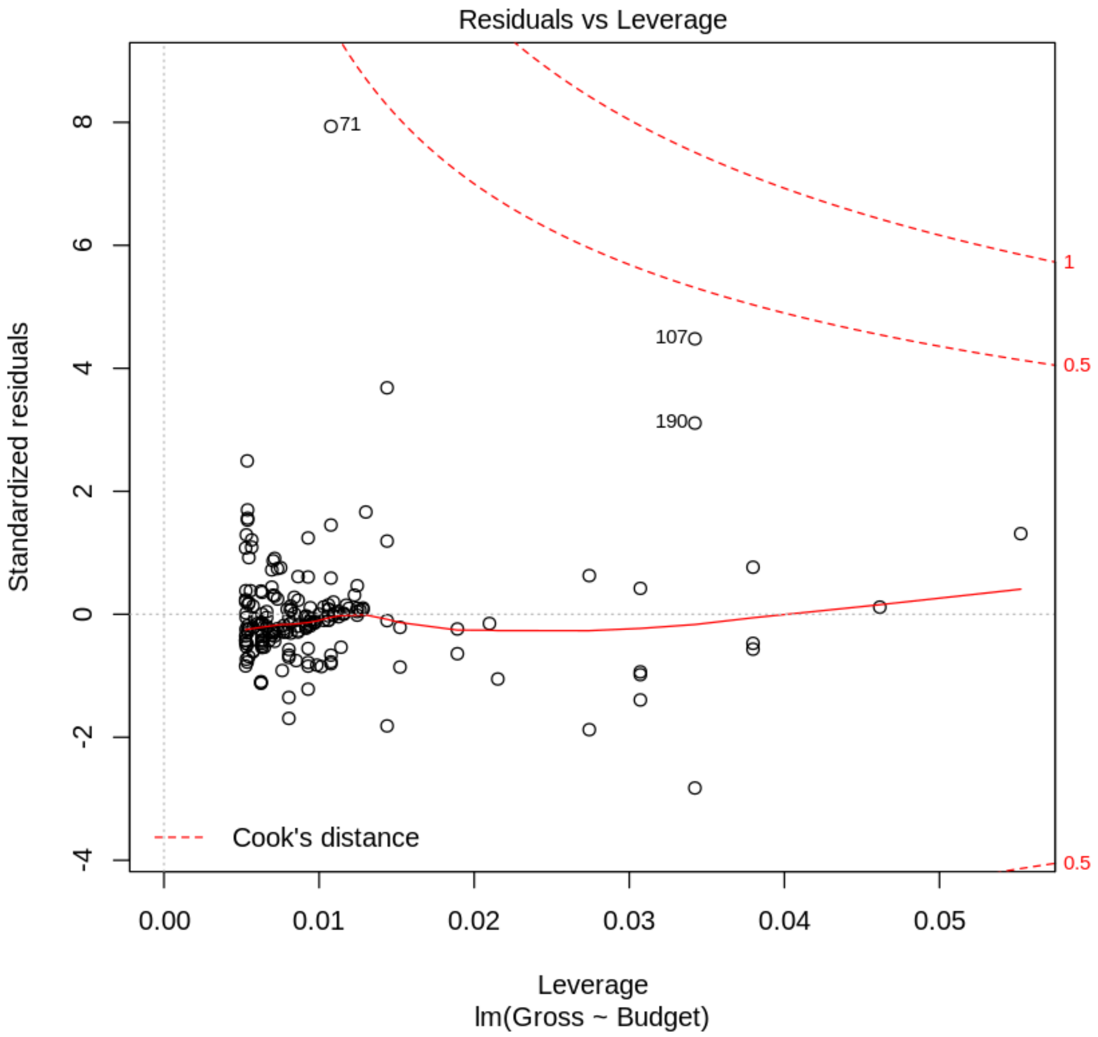
```
                    Movie          Gross              Budget
1920London             :  1   Min.    :  0.63   Min.    :  4.00
2 States\xa0           :  1   1st Qu.:  9.25   1st Qu.: 19.00
24(Tamil,Telugu)       :  1   Median : 29.38   Median : 34.50
Aashiqui 2             :  1   Mean   : 53.39   Mean   : 45.25
AeDilHainMushkil\xa0:  1   3rd Qu.: 70.42   3rd Qu.: 70.00
AGentleman             :  1   Max.   :500.75   Max.   :150.00
(Other)                :184
```

[50]:
```
plot(bolly.lm)
```




Normal Q-Q
Standardized residuals
-2
0
2
4
6
8
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(Gross ~ Budget)
71
107
117

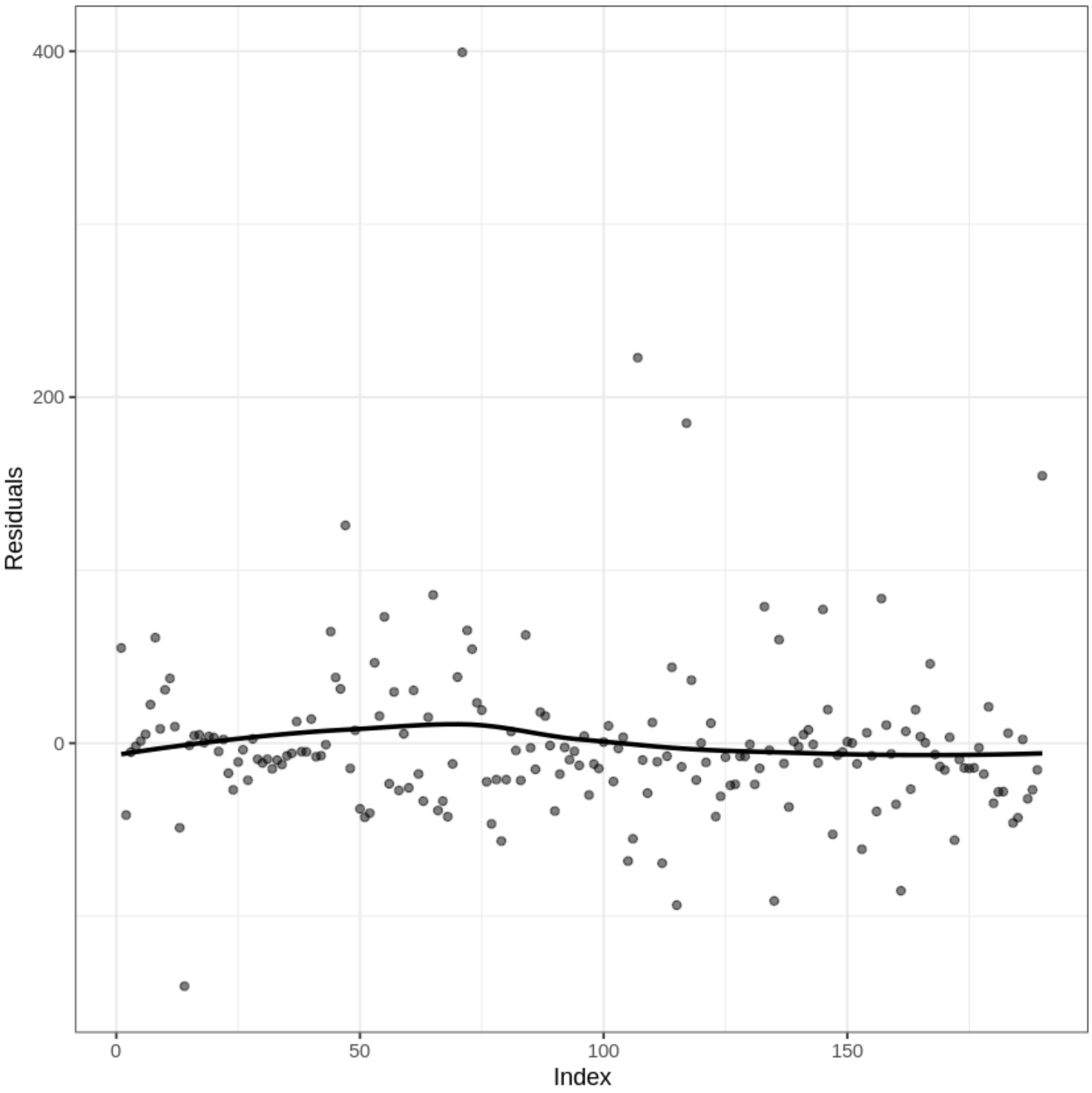

Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
0
50
100
150
200
71
107
117
Fitted values
lm(Gross ~ Budget)



[64]:

```
df.diagnostics.bolly = data.frame(yhat = fitted(bolly.lm), res = resid(bolly.lm), y=bollywood$Gross)

ggplot(df.diagnostics.bolly, aes(x = 1:length(sort(bollywood$Gross)), y = res)) +
 geom_point(alpha = 0.5) +
 xlab("Index") +
 geom_smooth(se = F, col = "black") +
 ylab("Residuals") +
 theme_bw()
```

```
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



[58]:
```
df.diagnostics.bolly[1,'Gross']
```

```
NULL
```

[ ]:

[ ]:

[ ]:

Looking at the plots we seem to have independence between error terms (residuals vs index plot). However, looking at the QQ plot and the Fitted values vs residuals plot we can see there are issue with constant variance. As the fitted values become largert, the residuals do as well.

3. (b) Transformations Notice that the Residuals vs. Fitted Values plot has a 'trumpet" shape to it, the points have a greater spread as the Fitted value increases. This means that there is not a constant variance, which violates the homoskedasticity assumption.

So how do we address this? Sometimes transforming the predictors or response can help stabilize the variance. Experiment with transfomrations on `Budget` and/or `Gross` so that, in the transformed scale, the relationship is approximately linear with a constant variance. Limit your transformations to square root, logarithms and exponentiation.

Note: There may be multiple transformations that fix this violation and give similar results. For the purposes of this problem, the transformed model doesn't have the be the "best" model, so long as it maintains both the linearity and homoskedasticity assumptions.

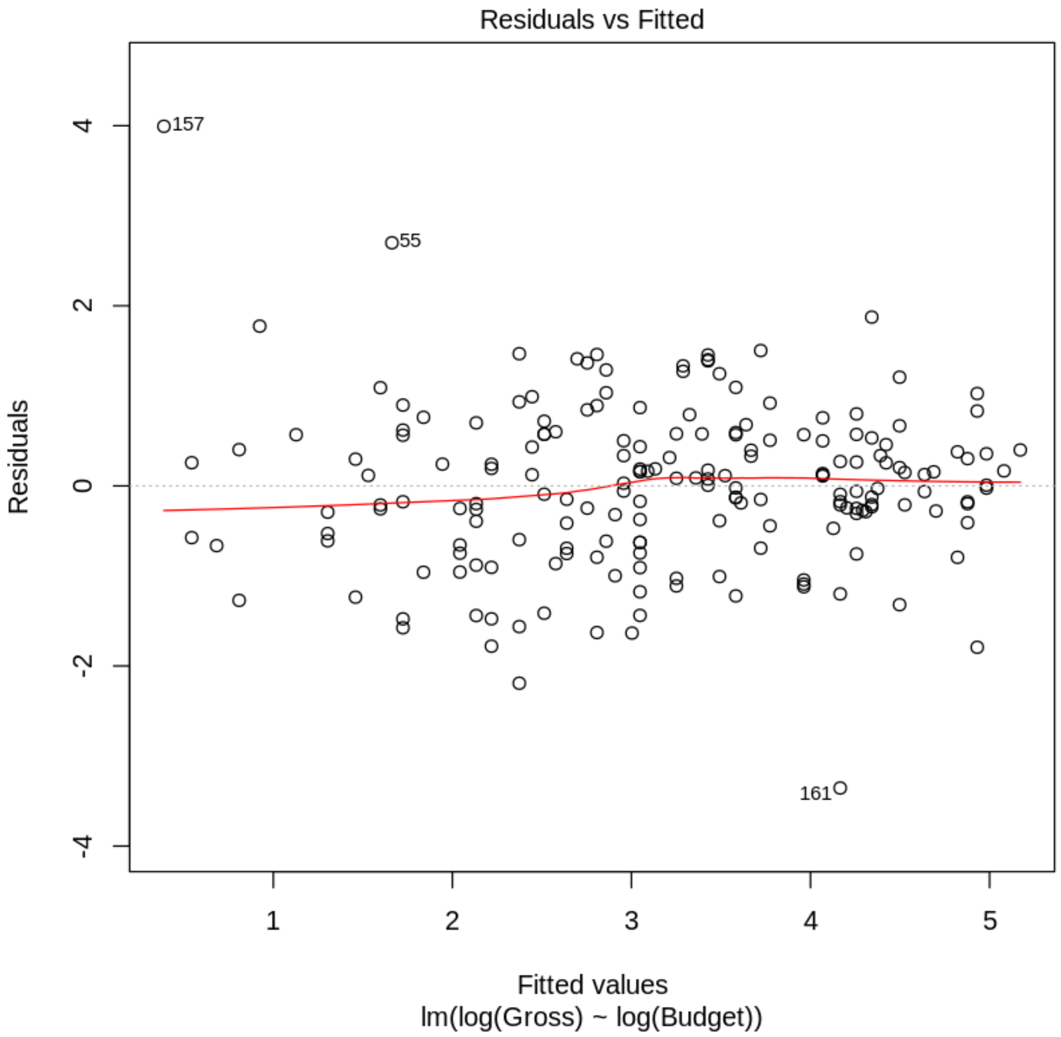
[88]:

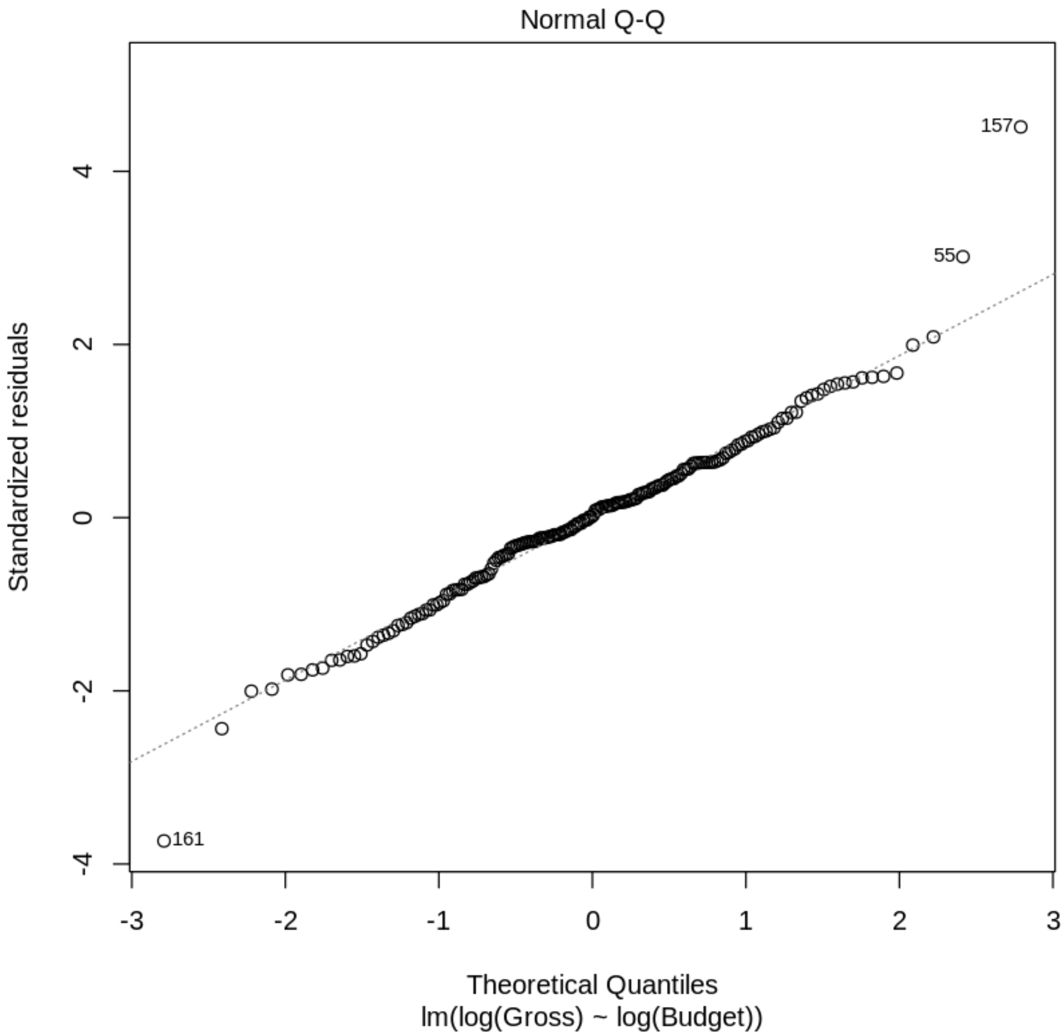
```
# Your Code Here

bolly.lm2 = lm(log(Gross)~log(Budget), data=bollywood)
summary(bolly.lm2)
```

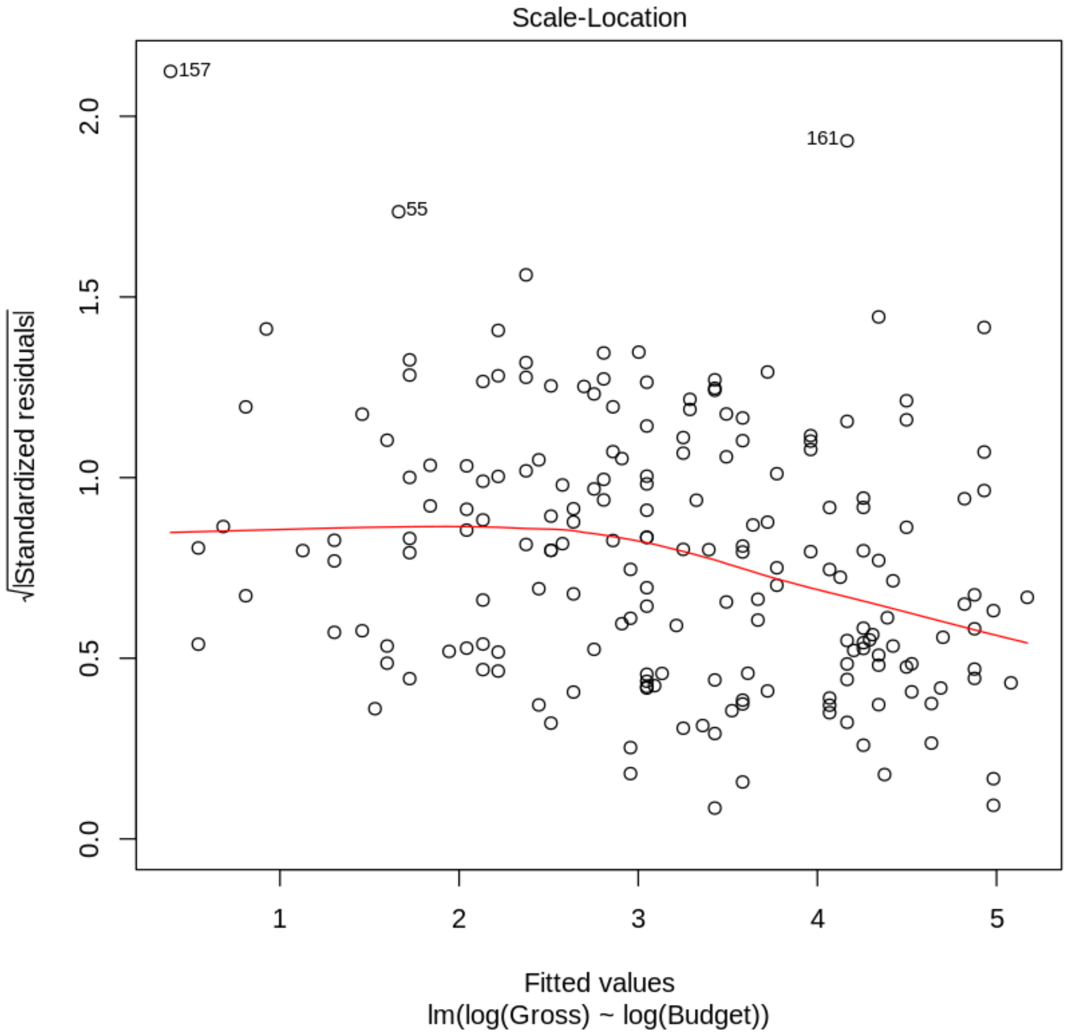
```
Call:
lm(formula = log(Gross) ~ log(Budget), data = bollywood)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.3549 -0.5634  0.0186  0.5664  3.9930 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.44023    0.28410  -5.069 9.51e-07 ***
log(Budget)  1.31955    0.07887  16.730  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9029 on 188 degrees of freedom
Multiple R-squared:  0.5982,Adjusted R-squared:  0.5961 
F-statistic: 279.9 on 1 and 188 DF,  p-value: < 2.2e-16
```

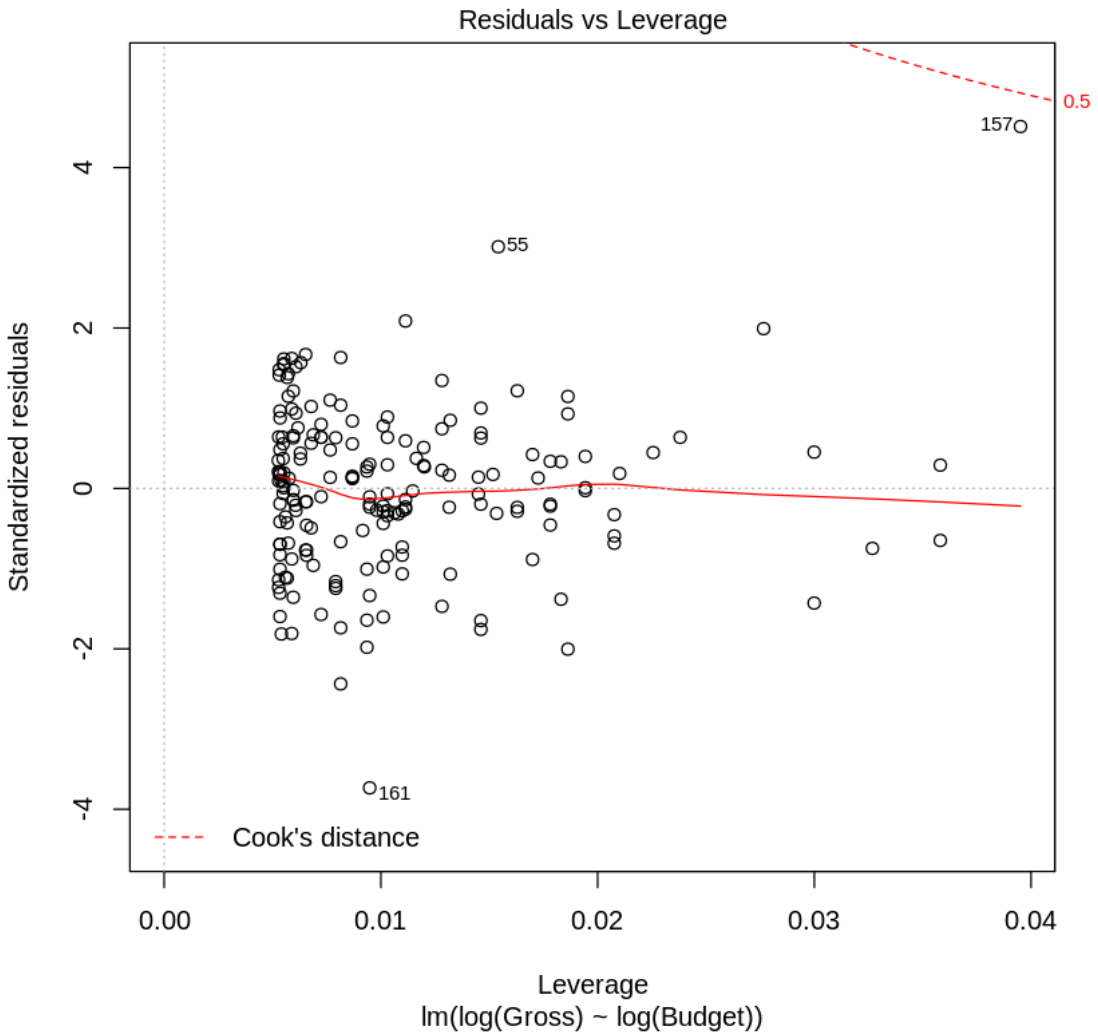
[87]:

```
plot(bolly.lm2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(log(Gross) ~ log(Budget))
4
2
0
-2
-4
1
2
3
4
5
157
55
161


Normal Q-Q
Standardized residuals
-4
-2
0
2
4
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(log(Gross) ~ log(Budget))
157
55
161


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
157
161
55
1
2
3
4
5
Fitted values
lm(log(Gross) ~ log(Budget))



3. (c) Interpreting Your Transformation You've fixed the nonconstant variance problem! Hurray! But now we have a transformed model, and it will have a different interpretation than a normal linear regression model. Write out the equation for your transformed model. Does this model have an interpretation similar to a standard linear model?

log(Gross) = -1.44023 + 1.31955 * log(Budget) + ε

This model has a different interpreation than before. The model can be interpretated as the following: A one percent increase in the Budget of the movie increases the Gross amount by 1.32%.

[ ]: