C1M5_peer_reviewed

June 7, 2023

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.

[1]:
```
# 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.

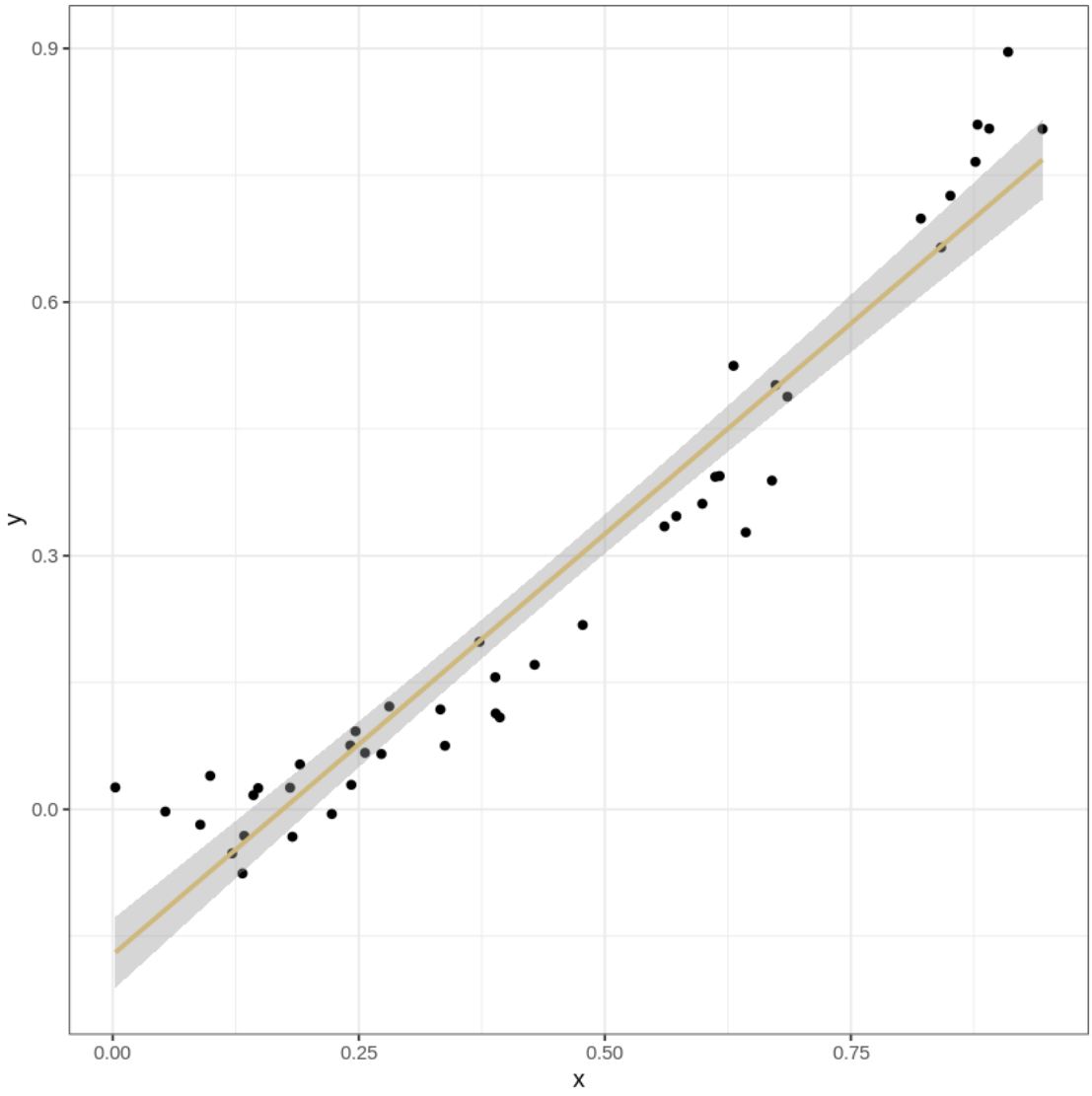
Then fit a linear model to these data and comment on where you can diagnose nonlinearity in the diagnostic plots.

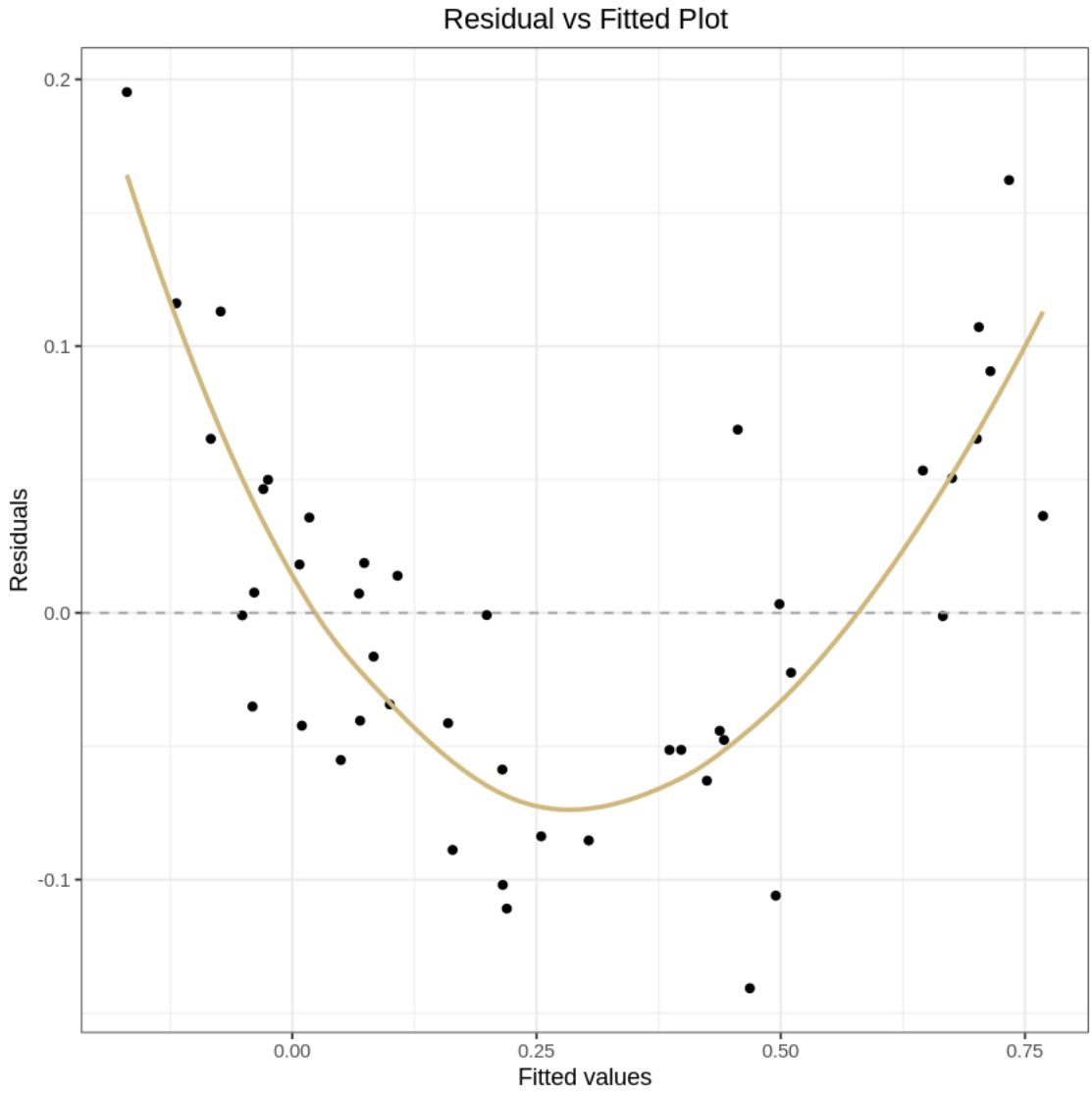
[2]:
```
# Your Code Here
set.seed(2016)
library(ggplot2)
# Simulation for violating linearity
n = 45; x = runif(n, 0, 1); y = 1 + x + rnorm(n,0,abs(x))
y = x^2 + rnorm(n, sd = 0.05)
```

```r
lmod = lm(y ~ x);

ggplot(mapping = aes(x = x, y = y)) +
    geom_point() +
    geom_smooth(method = "lm", col = "#CFB87C")+theme_bw()

#diagnostic plot
p1=ggplot(lmod, aes(.fitted, .resid))+geom_point()
p1= p1+stat_smooth(method="loess", col = "#CFB87C", se = FALSE)
p1=p1 + geom_hline(yintercept=0, col="#A2A4A3", linetype="dashed")
p1=p1+xlab("Fitted values")+ylab("Residuals")
    p1<-p1+ggtitle("Residual vs Fitted Plot")+theme_bw() +
    theme(plot.title = element_text(hjust=0.5))


p1
```

```
`geom_smooth()` using formula 'y ~ x'
```

```
`geom_smooth()` using formula 'y ~ x'
```


0.9
0.6
0.3
0.0
y
0.00
0.25
0.50
0.75
x



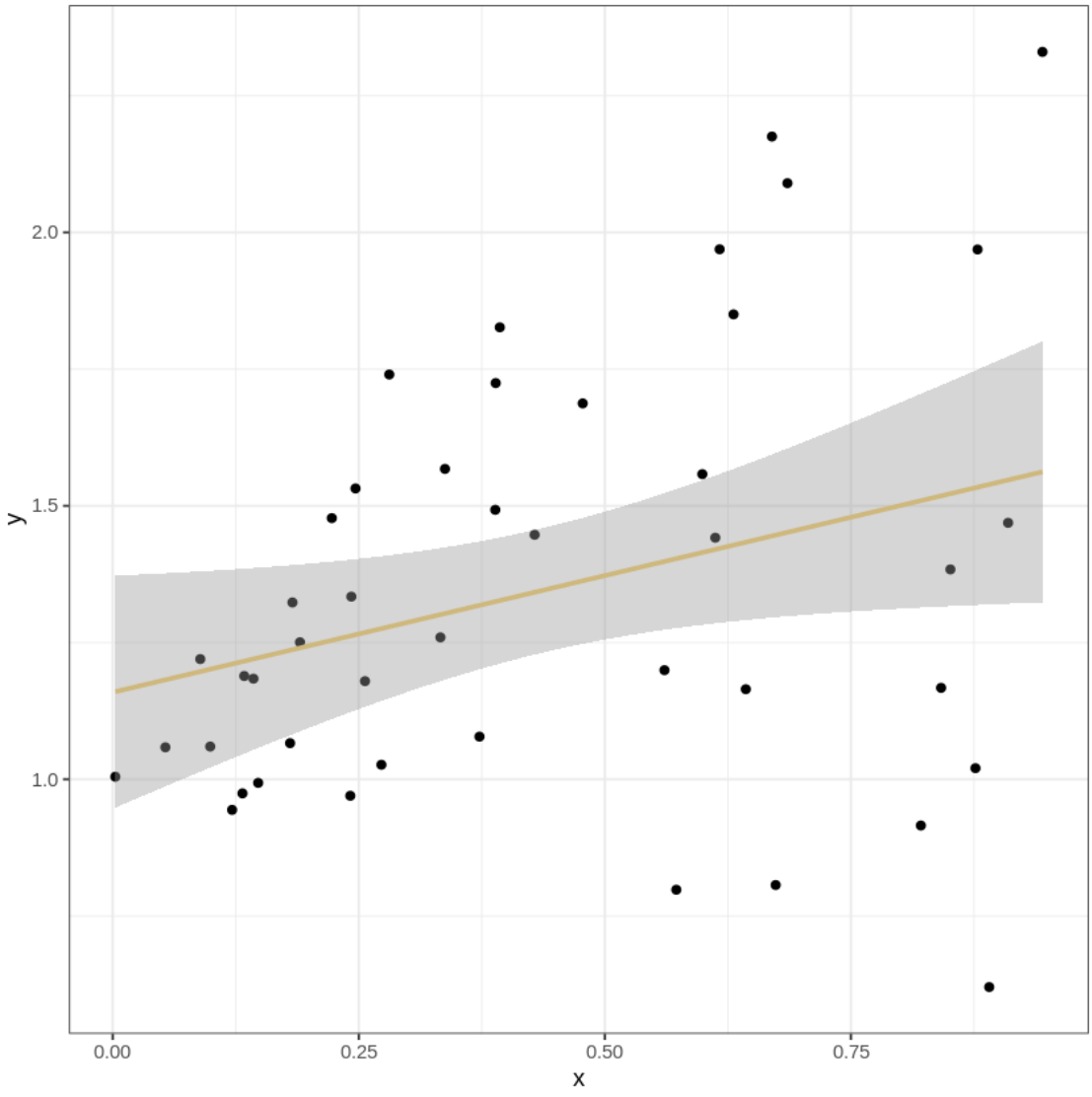
1. (b) Homoskedasticity Simulate another SLR dataset that violates the constant variance assumption, but maintains the other assumptions. Then fit a linear model to these data and comment on where you can diagnose non-constant variance in the diagnostic plots.

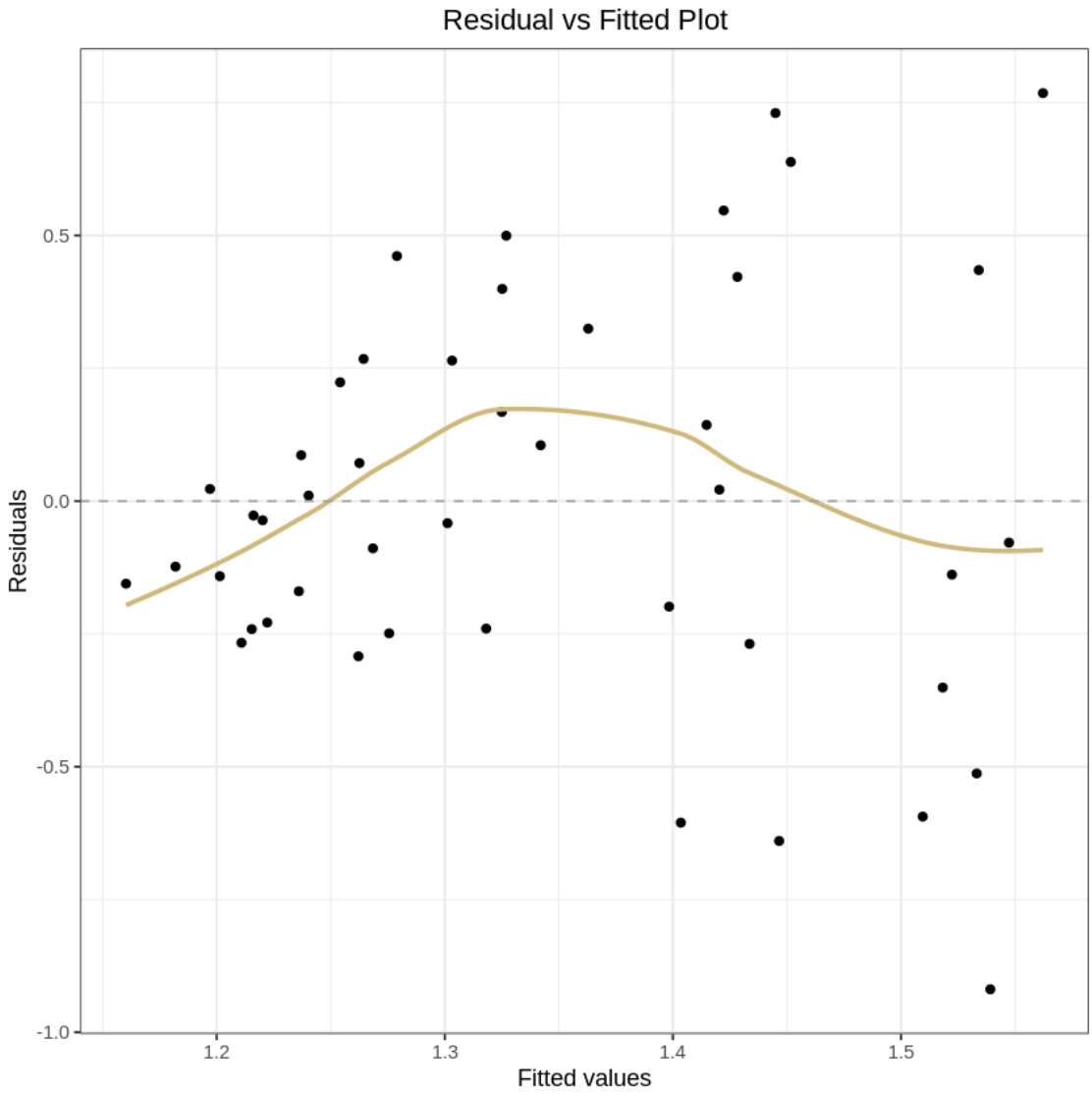
[3]:
```
# Your Code Here
y = 1 + x + rnorm(n,0,abs(x))
lmod = lm(y ~ x)

ggplot(mapping = aes(x = x, y = y)) +
    geom_point() +
    geom_smooth(method = "lm", col = "#CFB87C")+theme_bw()
```

```r
#diagnostic plot
p1=ggplot(lmod, aes(.fitted, .resid))+geom_point()
p1= p1+stat_smooth(method="loess", col = "#CFB87C", se = FALSE)
p1=p1 + geom_hline(yintercept=0, col="#A2A4A3", linetype="dashed")
p1=p1+xlab("Fitted values")+ylab("Residuals")
    p1<-p1+ggtitle("Residual vs Fitted Plot")+theme_bw() +
    theme(plot.title = element_text(hjust=0.5))


p1
```

```
`geom_smooth()` using formula 'y ~ x'
```

```
`geom_smooth()` using formula 'y ~ x'
```





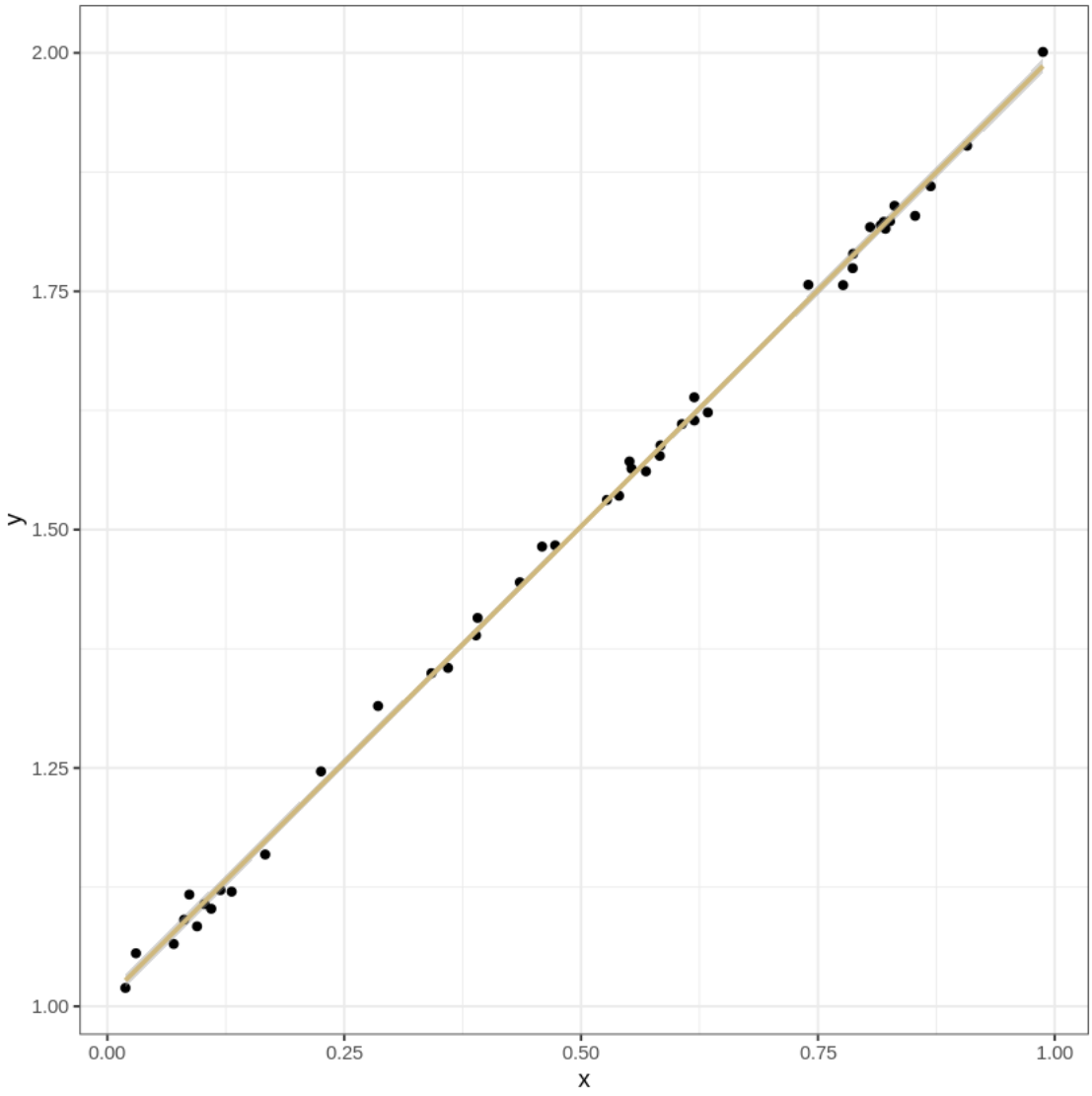
1. (c) Independent Errors Repeat the above process with simulated data that violates the independent errors assumption.

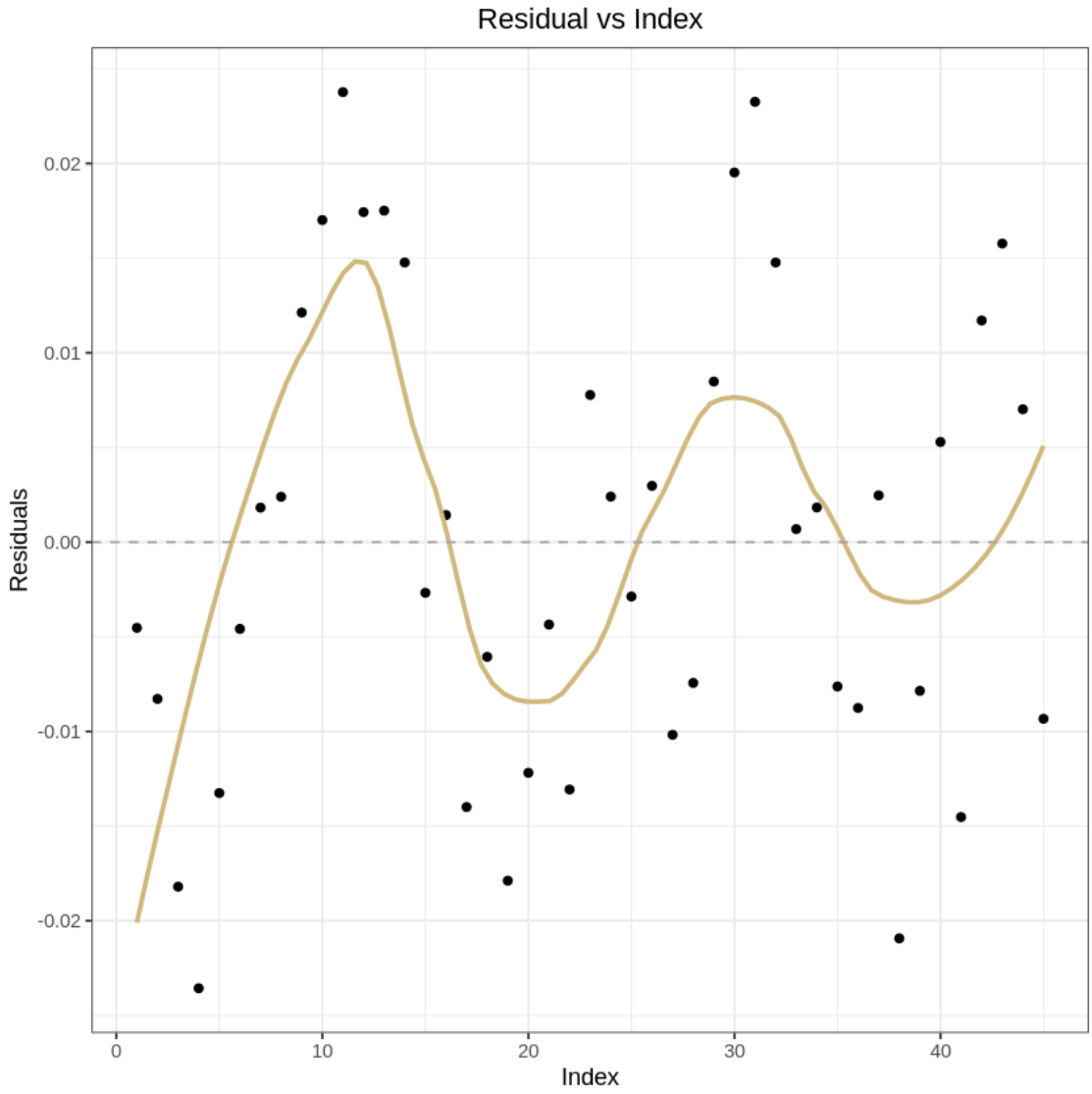
[4]:
```
# Your Code Here
set.seed(999)
n = 45; x = runif(n, 0, 1);

e = matrix(NA, nrow = n)

rho = 0.7
```

```
e[1] = 0
for (i in 2:n){
    e[i] = rho*e[i-1] + rnorm(1,0,0.01)
}

y = 1 + x + e
lmod = lm(y ~ x);


ggplot(mapping = aes(x = x, y = y)) +
    geom_point() +
    geom_smooth(method = "lm", col = "#CFB87C")+theme_bw()

#diagnostic plot
p1=ggplot(lmod, aes(1:n, .resid))+geom_point()
p1= p1+stat_smooth(method="loess", col = "#CFB87C", se = FALSE, span = 0.5)
p1=p1 + geom_hline(yintercept=0, col="#A2A4A3", linetype="dashed")
p1=p1+xlab("Index")+ylab("Residuals")
    p1<-p1+ggtitle("Residual vs Index")+theme_bw() +
    theme(plot.title = element_text(hjust=0.5))


p1
```

```
`geom_smooth()` using formula 'y ~ x'
```

```
`geom_smooth()` using formula 'y ~ x'
```


2.00
1.75
1.50
1.25
1.00
y
0.00
0.25
0.50
0.75
1.00
x



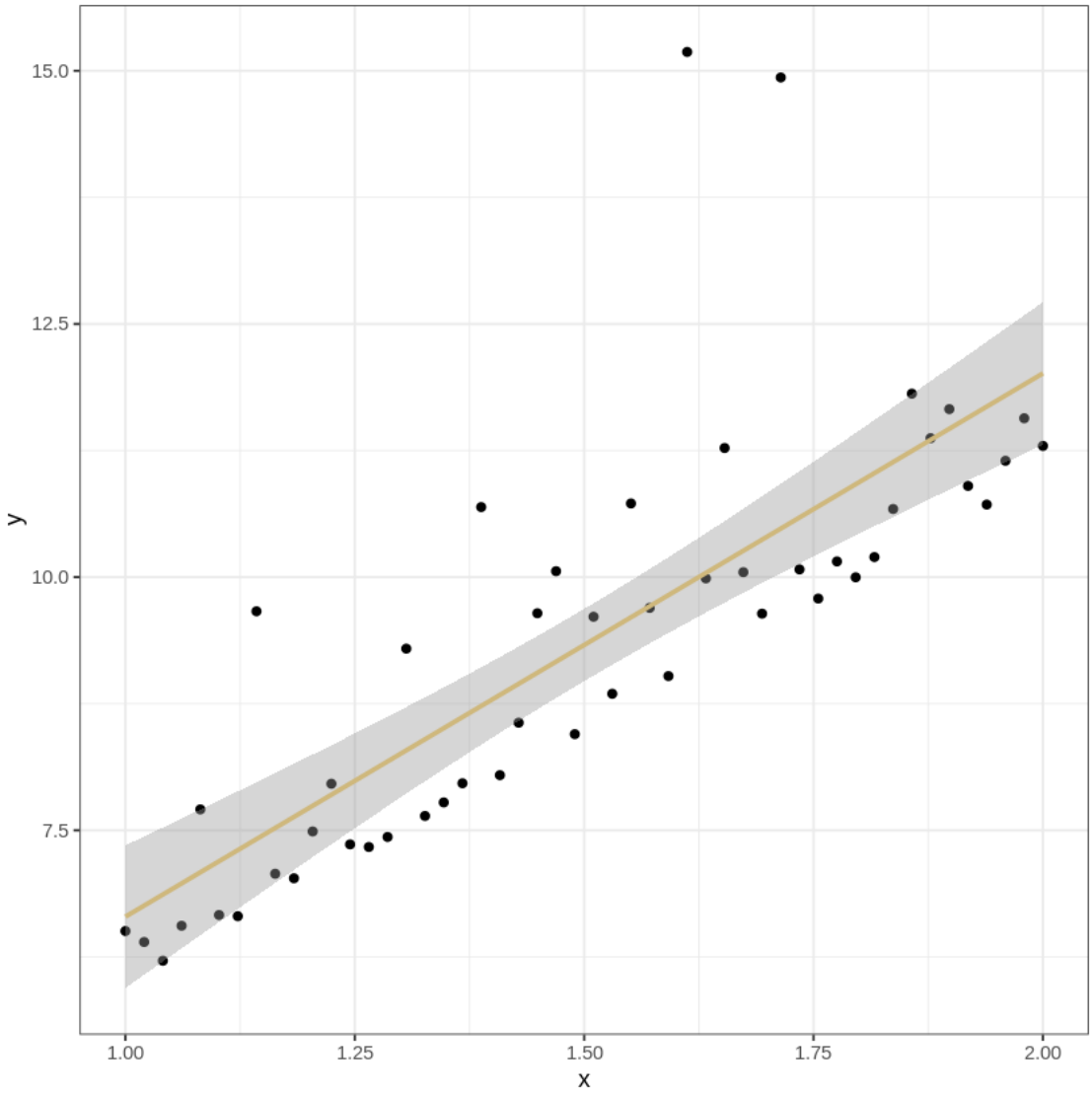
1. (d) Normally Distributed Errors Only one more to go! Repeat the process again but simulate the data with non-normal errors.

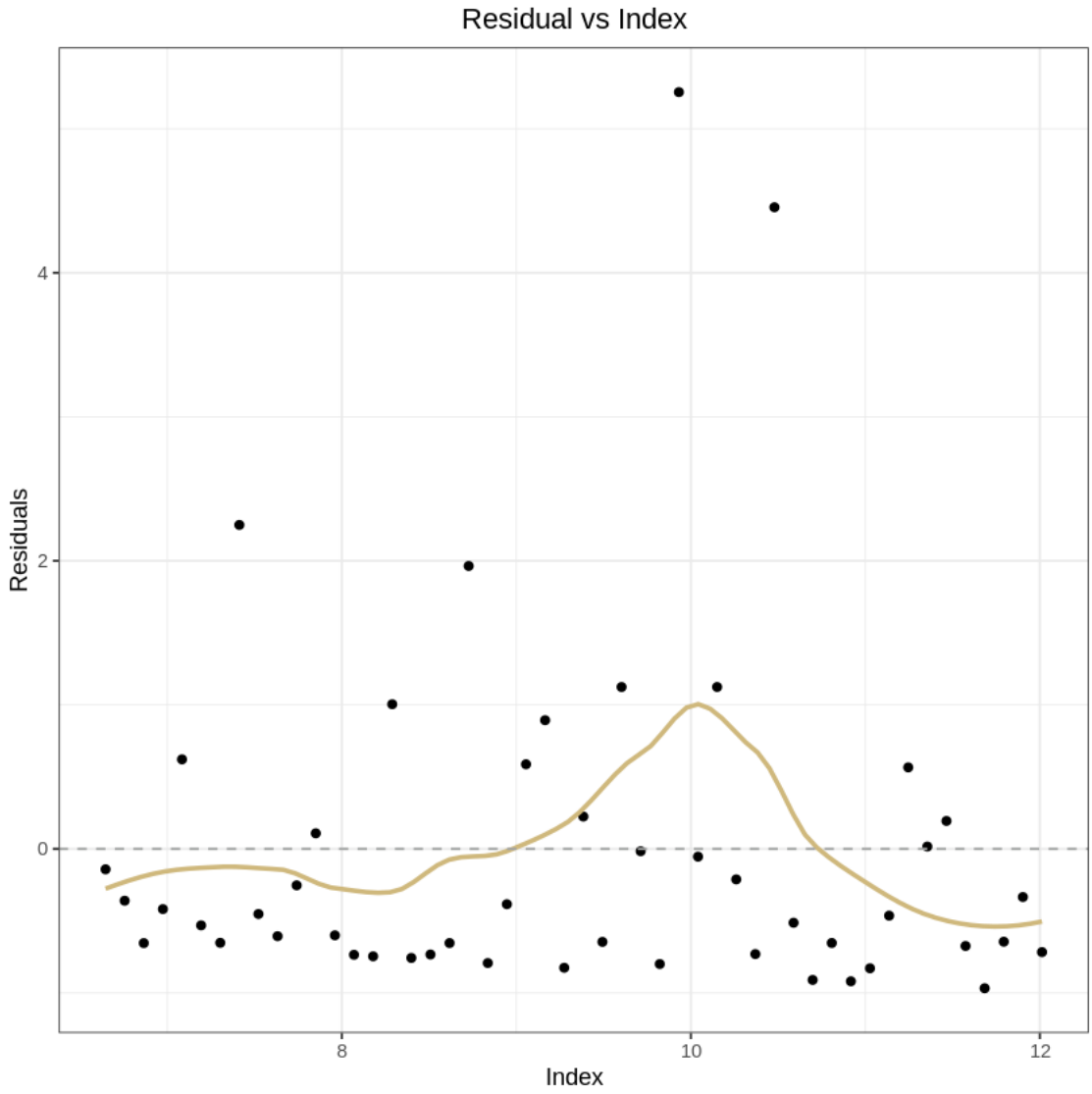
[5]:
```
# Your Code Here
n = 50; x = seq(1,2, length.out = n); b0 = 1; b1 = 5; e = rgamma(n,0.5, 0.5)
y = b0 + b1*x + e

lmod = lm(y ~ x)

ggplot(mapping = aes(x = x, y = y)) +
    geom_point() +
    geom_smooth(method = "lm", col = "#CFB87C")+theme_bw()
```

```r
#diagnostic plot
p1=ggplot(lmod, aes(.fitted, .resid))+geom_point()
p1= p1+stat_smooth(method="loess", col = "#CFB87C", se = FALSE, span = 0.5)
p1=p1 + geom_hline(yintercept=0, col="#A2A4A3", linetype="dashed")
p1=p1+xlab("Index")+ylab("Residuals")
    p1<-p1+ggtitle("Residual vs Index")+theme_bw() +
    theme(plot.title = element_text(hjust=0.5))


p1
```

```
`geom_smooth()` using formula 'y ~ x'
```

```
`geom_smooth()` using formula 'y ~ x'
```


15.0
12.5
10.0
7.5
y
1.00
1.25
1.50
1.75
2.00
x



2 Problem 2: Hats for Sale

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

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

The goal of this question is to use the hat matrix to prove that the fitted values, $\widehat{\mathbf{Y}}$, and the residuals, $\widehat{\boldsymbol{\varepsilon}}$, 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. Yˆ=Xˆ=X(XTX)−1XTY=HY

2. (b) Show that H is symmetric: $H = H^T$.
HT=(X(XTX)−1XT)T=(XT)T[(XTX)−1]TXT=X[(XTX)T]−1XT=X(XTX)−1XT=H

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

Yˆ T ˆ=(HY)T(I−H)Y=YTHT(In−H)Y=YTH(In−H)Y=YT0nY=0.

2.(e) Why is this result important in the practical use of linear regression? if the linear regression assumptions are met, then there should be no correlation between the fitted values, Yˆ , and the residuals, ˆ .

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.

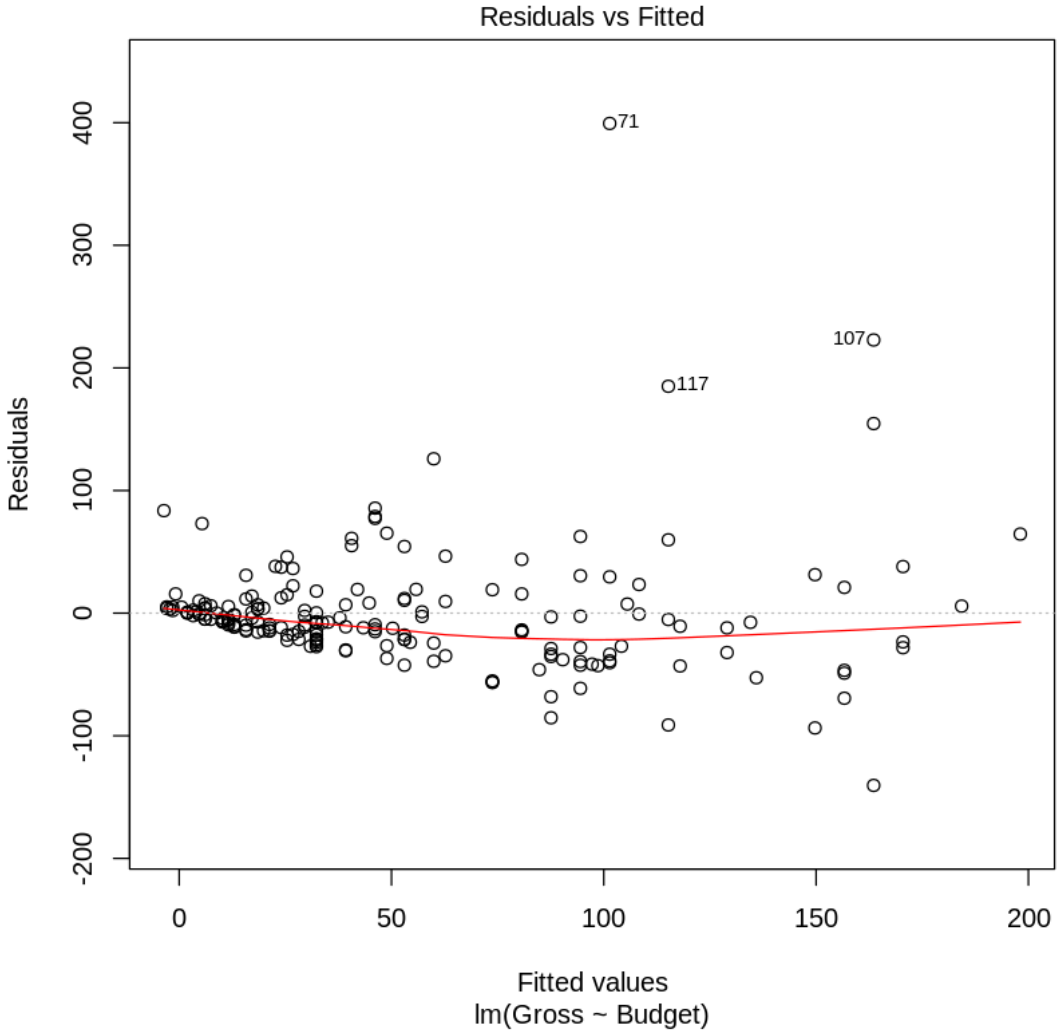
Specify whether each of the four regression model assumptions are being violated.

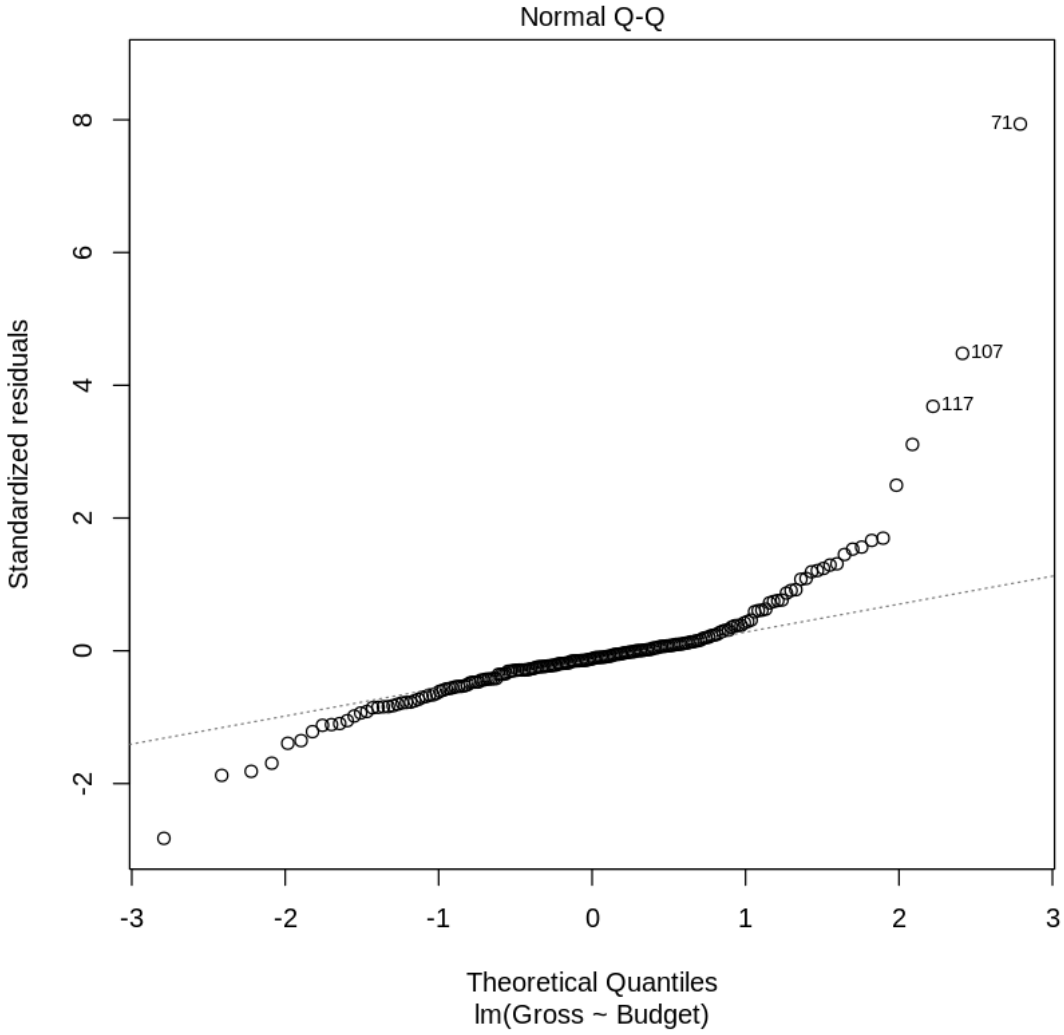
Data Source: http://www.bollymoviereviewz.com

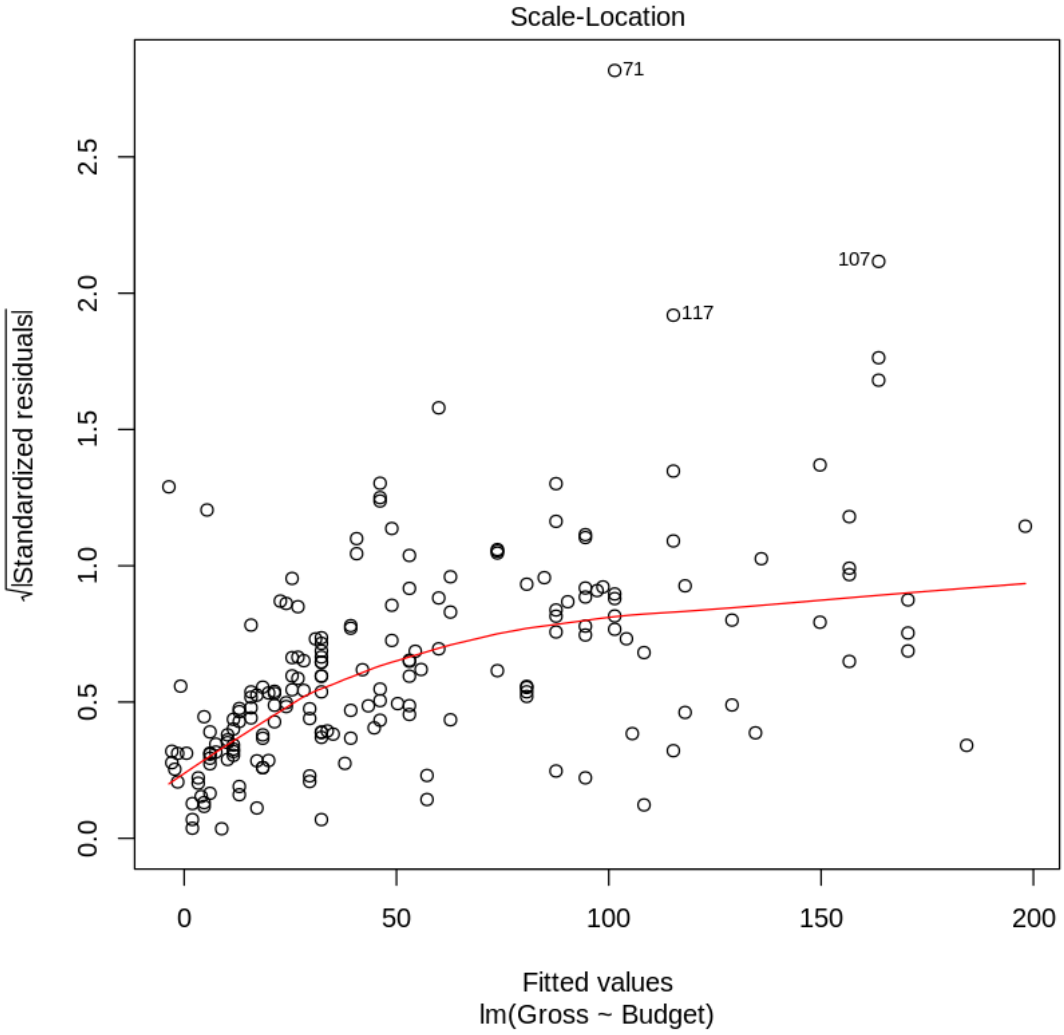
[6]:
```
# Load the data
bollywood = read.csv("bollywood_boxoffice.csv")
summary(bollywood)

# Your Code Here
lm_bollywood = lm(Gross ~ Budget, data = bollywood)

plot(lm_bollywood)


p1=ggplot(lm_bollywood, aes(1:dim(bollywood)[1], .resid))+geom_point()
p1= p1+stat_smooth(method="loess", col = "#CFB87C", se = FALSE, span = 0.3)
```

```
p1=p1 + geom_hline(yintercept=0, col="#A2A4A3", linetype="dashed")
p1=p1+xlab("Index")+ylab("Residuals")
    p1<-p1+ggtitle("Residual vs Index")+theme_bw() +
    theme(plot.title = element_text(hjust=0.5))

p1

n = dim(bollywood)[1];
x = head(resid(lm_bollywood), n-1)
y = tail(resid(lm_bollywood), n-1)
cor(x,y)
srp = data.frame(x,y)
ggplot(srp, aes(x = x, y = y)) +
    geom_point() +
    geom_vline(xintercept = 0) +
    geom_hline(yintercept = 0) +
    xlab(expression(hat(epsilon)[i])) +
    ylab(expression(hat(epsilon)[i+1])) +
    ggtitle("Successive Residual Plot") +
    theme_bw() +
    theme(plot.title = element_text(hjust = 0.5))
```

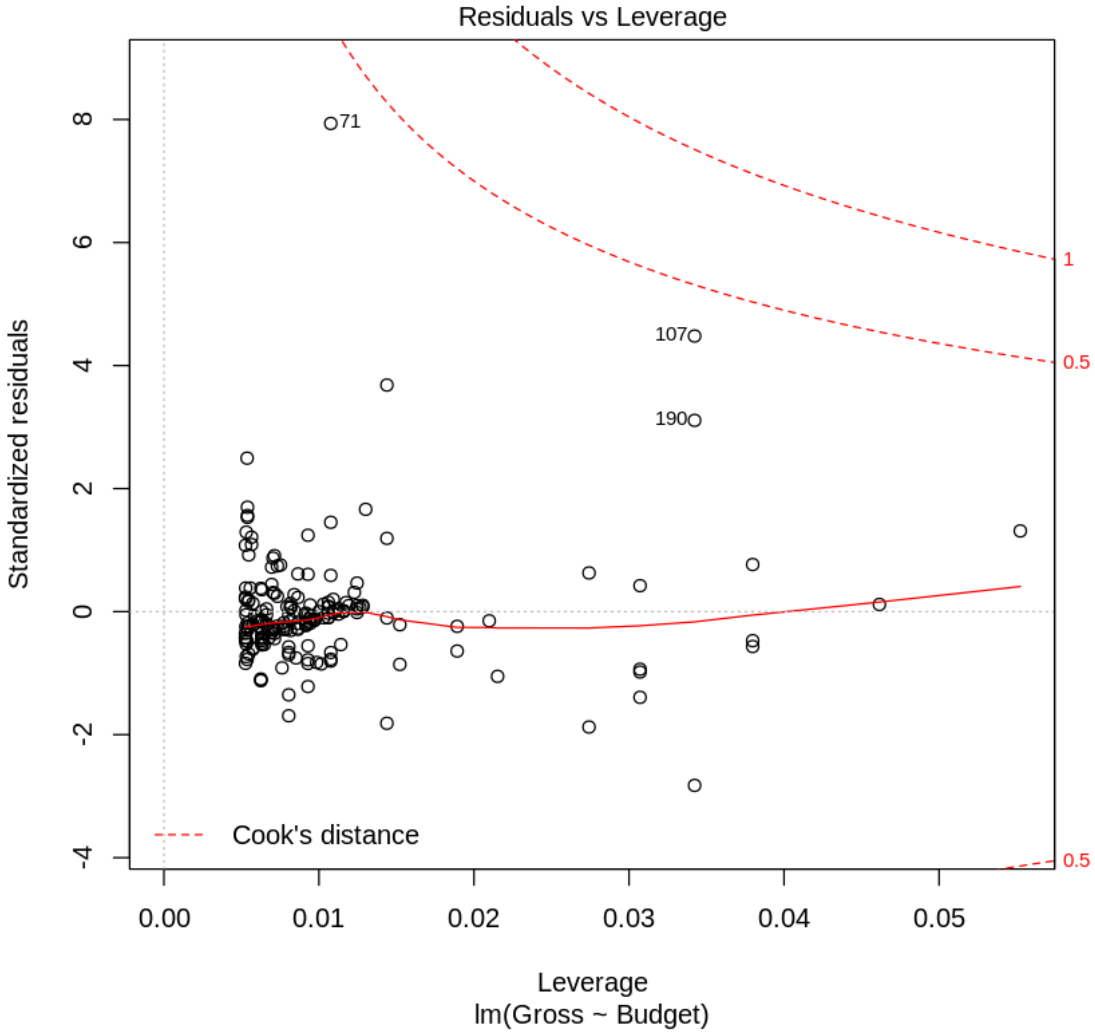
```
                 Movie         Gross            Budget
1920London          :  1   Min.   :  0.63   Min.   :  4.00
2 States\xa0        :  1   1st Qu.:  9.25   1st Qu.: 19.00
24(Tamil,Telugu)    :  1   Median : 29.38   Median : 34.50
Aashiqui 2          :  1   Mean   : 53.39   Mean   : 45.25
AeDilHainMushkil\xa0: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


Residuals vs Fitted
Residuals
Fitted values
lm(Gross ~ Budget)
71
107
117
400
300
200
100
0
-100
-200
0
50
100
150
200

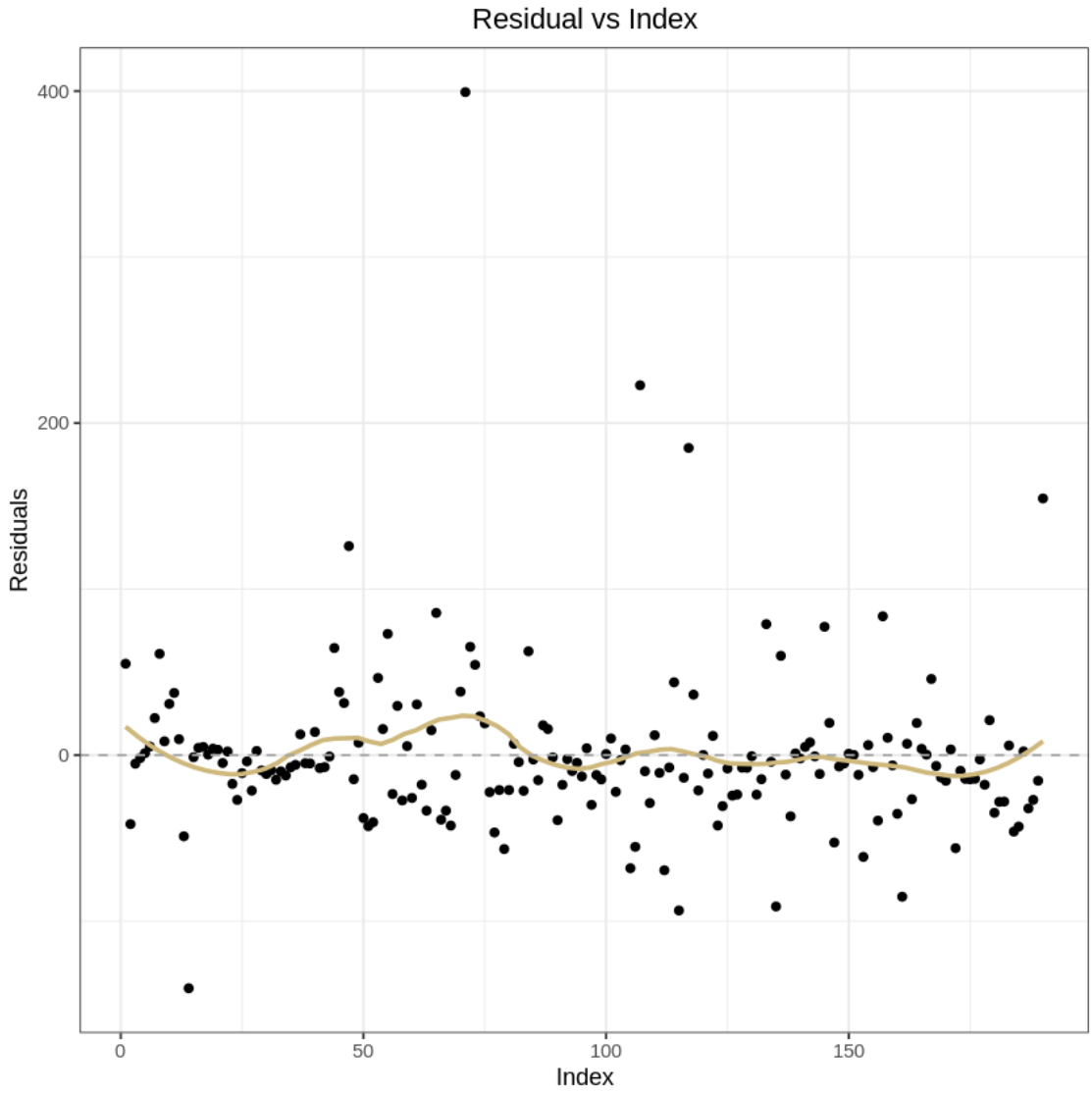

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Gross ~ Budget)
71
107
117
-3
-2
-1
0
1
2
3
-2
0
2
4
6
8

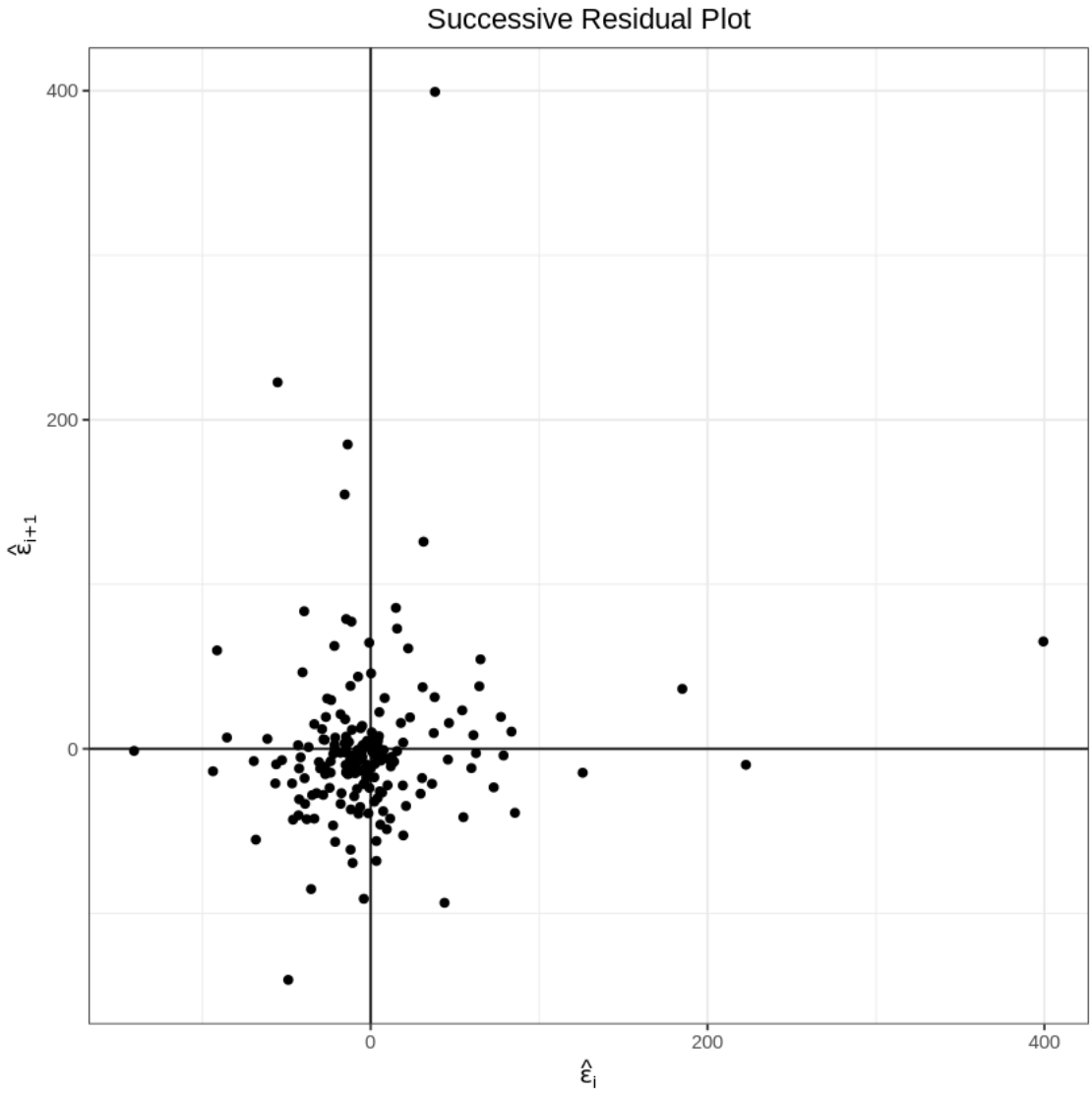



```
`geom_smooth()` using formula 'y ~ x'
```



0.111594190315052


Residual vs Index
Residuals
400
200
0
0
50
100
150
Index



In the analysis described, several assumptions of linear regression are being evaluated based on diagnostic plots. Here is a summary of the findings:

Linearity: The residual vs fitted plot exhibits a downward trend, indicating a possible violation of the linearity assumption. This suggests that the relationship between the predictors and the response may not be adequately captured by a linear model.

Independence: The residual vs index plot shows little structure, and the successive residuals plot indicates a weak or no correlation between successive error terms. Therefore, there is no evidence of dependence among the error terms in the current ordering of the data. However, it is acknowledged that different orderings of the data may reveal a dependence structure.

Constant variance: There is some evidence, although weak, of non-constant variance. The residual vs fitted plot displays more variability for larger fitted values compared to smaller fitted values. This violates the assumption of constant variance, also known as homoscedasticity, where the spread of

the residuals should be consistent across all levels of the predictors.

Normality: The QQ-plot demonstrates a clear deviation from normality. However, it is important to note that this deviation could be a result of the violations of linearity and constant variance assumptions. By addressing and remedying these issues, it is possible to restore the normality assumption in the model.

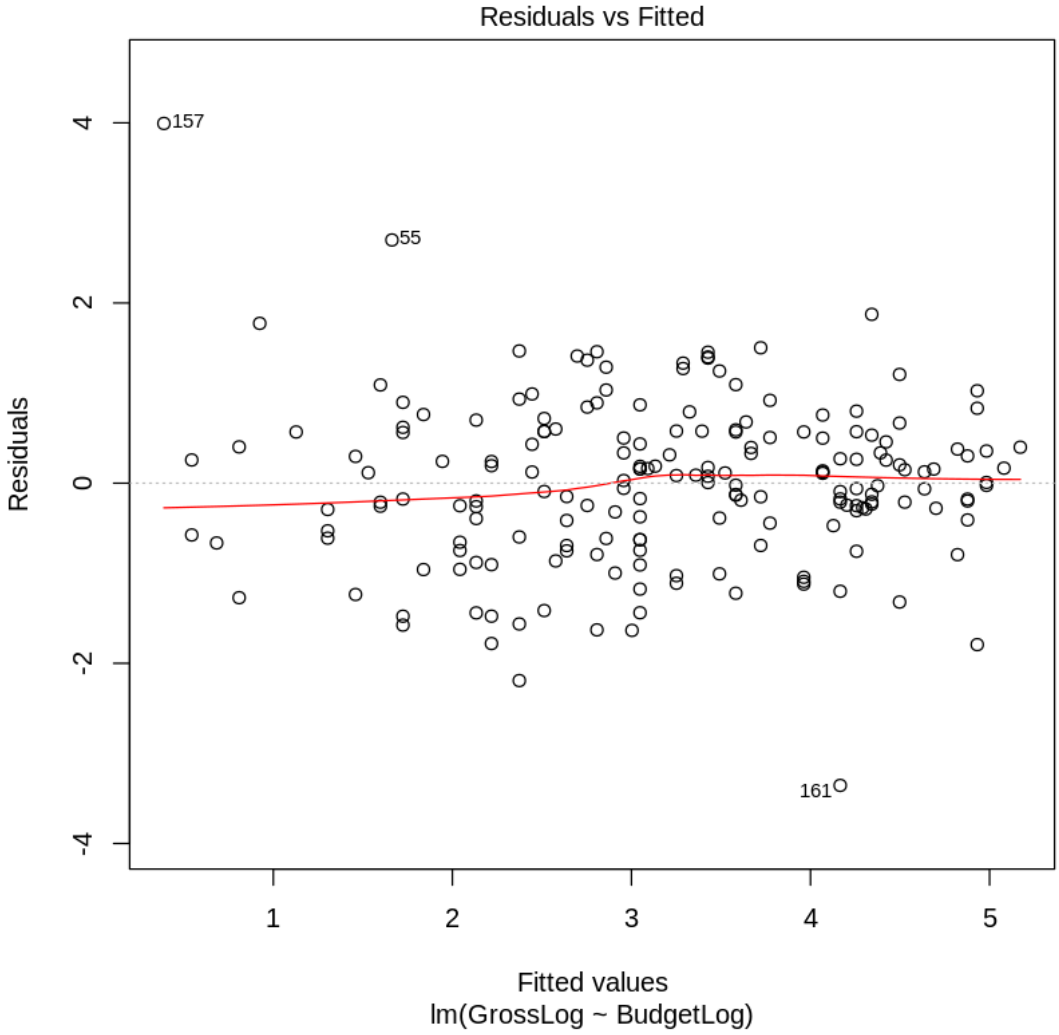
In summary, based on the diagnostic plots, there are indications of potential violations of the linearity, constant variance, and normality assumptions. These findings suggest the need for further investigation and potentially adjusting the model to address these issues. By addressing these violations, the model can be improved to better meet the assumptions of linear regression.

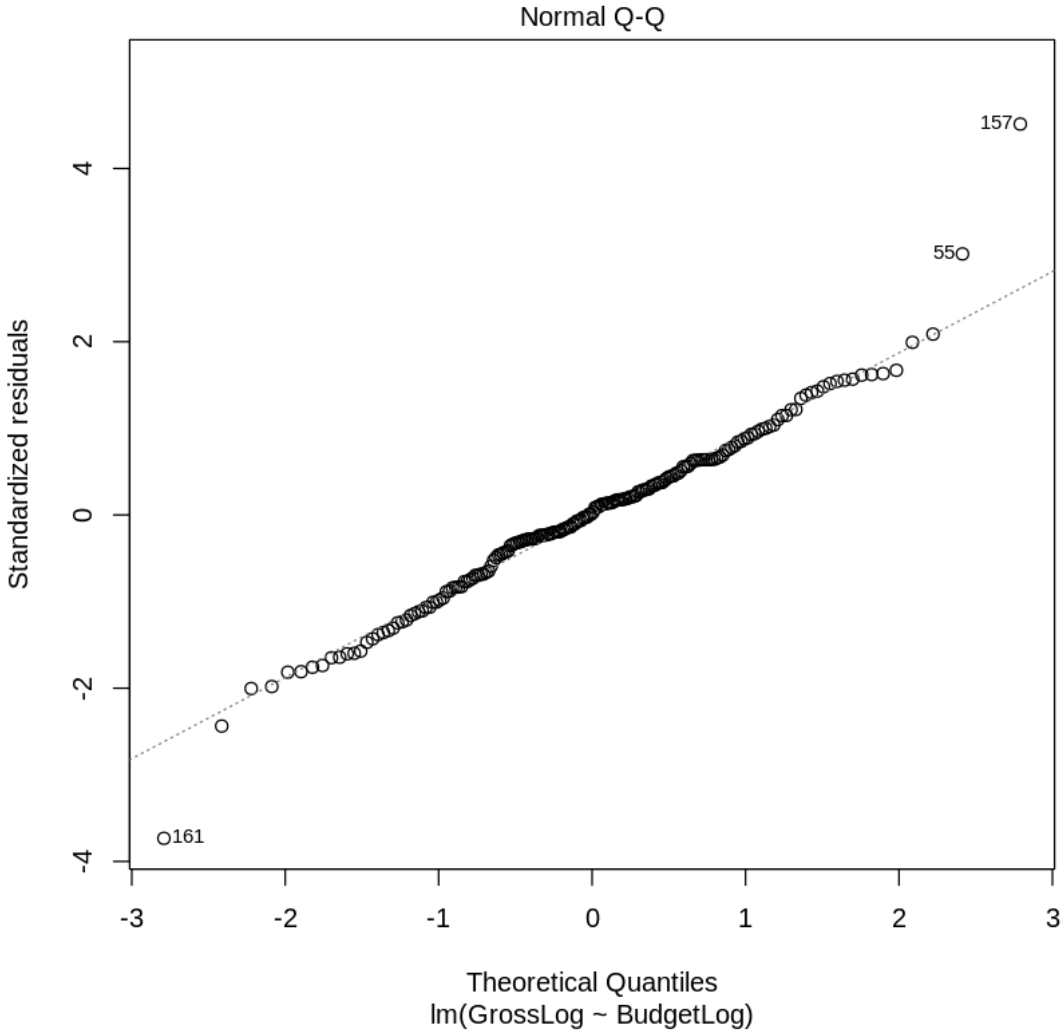
3. (b) Transformations Notice that the Residuals vs. Fitted Values plot has a 'trumpet" shape to it, the points have a greater spread as the Fitted value increases. This means that there is not a constant variance, which violates the homoskedasticity assumption.

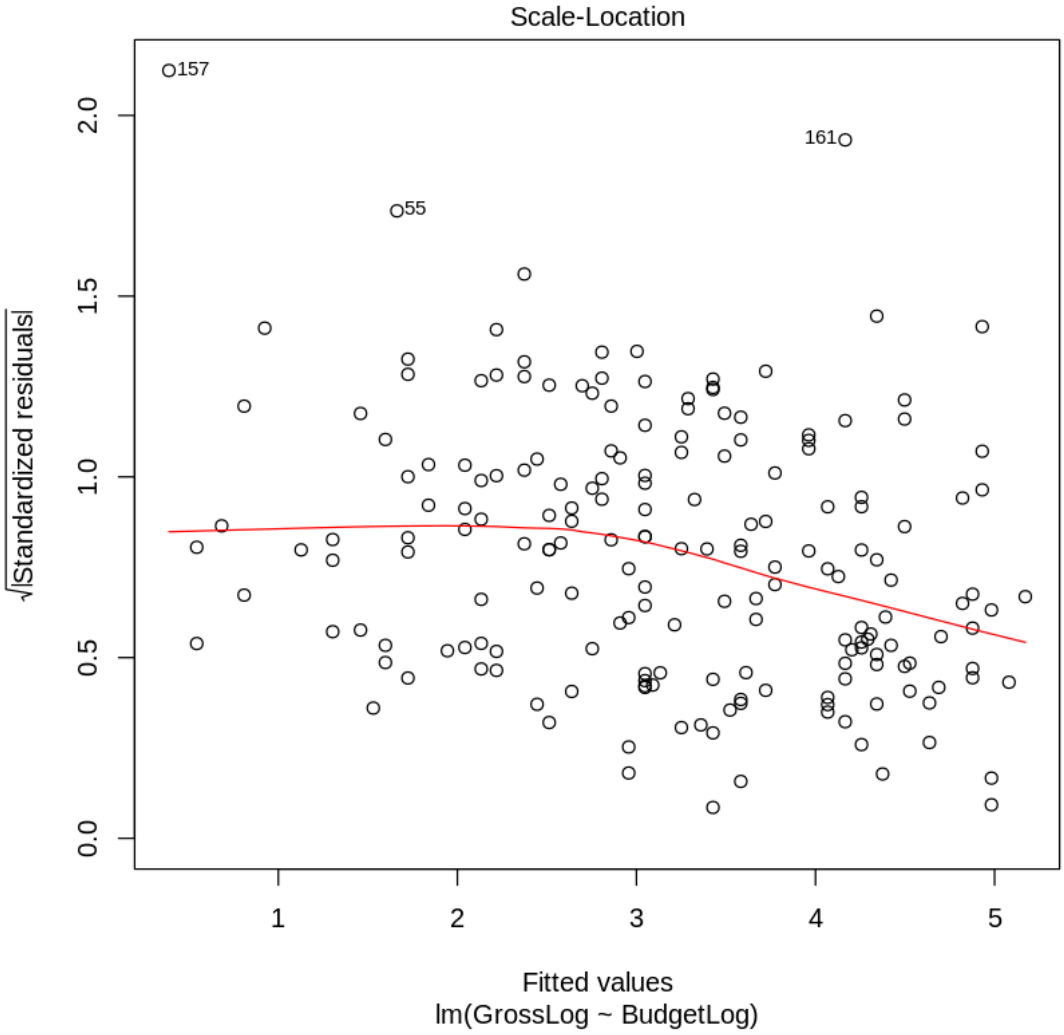
So how do we address this? Sometimes transforming the predictors or response can help stabilize the variance. Experiment with transfomrations on `Budget` and/or `Gross` so that, in the transformed scale, the relationship is approximately linear with a constant variance. Limit your transformations to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

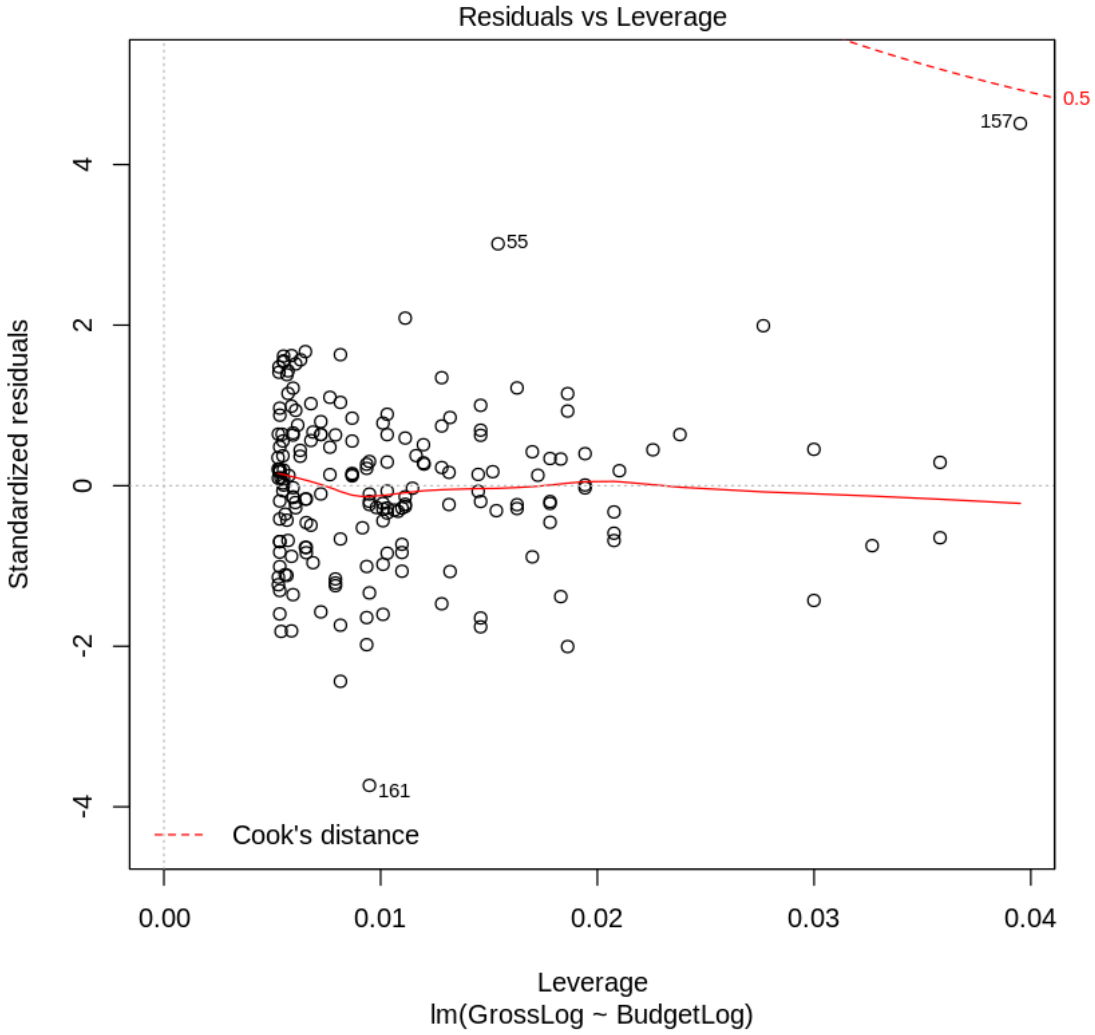
Note: There may be multiple transformations that fix this violation and give similar results. For the purposes of this problem, the transformed model doesn't have the be the "best" model, so long as it maintains both the linearity and homoskedasticity assumptions.

[7]:
```
# Your Code Here
bollywood$BudgetLog = log(bollywood$Budget)
bollywood$GrossLog = log(bollywood$Gross)
bollywood$BudgetSqrt = sqrt(bollywood$Budget)
bollywood$GrossSqrt = sqrt(bollywood$Gross)

lm_log_log = lm(GrossLog ~ BudgetLog, bollywood)
plot(lm_log_log)
```


Residuals vs Fitted
Residuals
Fitted values
lm(GrossLog ~ BudgetLog)
157
55
161
4
2
0
-2
-4
1
2
3
4
5


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(GrossLog ~ BudgetLog)
157
55
161
-4
-2
0
2
4
-3
-2
-1
0
1
2
3


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



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?