# Module 5: Peer Reviewed Assignment

### 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.

In [8]:

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

## 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.

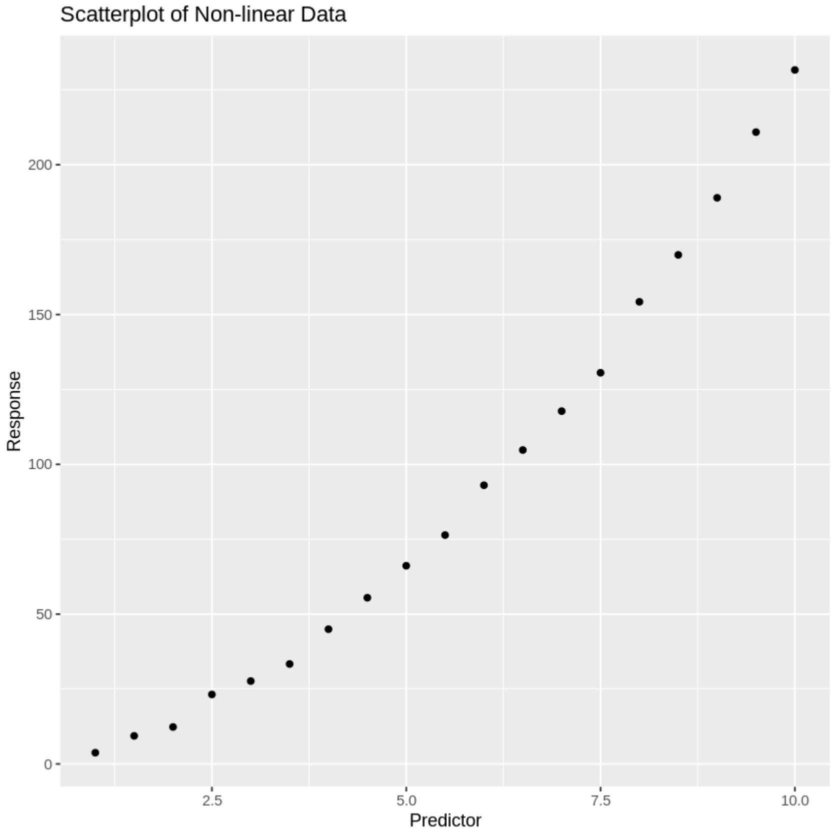
#### 1. (a) Linearity

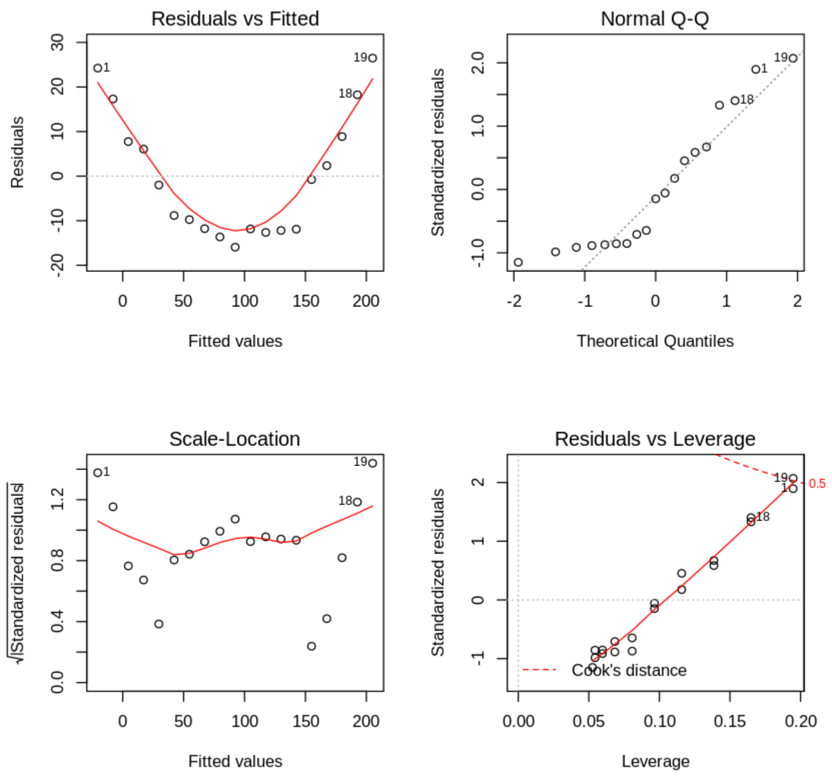
Generate SLR data that violates the linearity assumption, but maintains the other assumptions. Create a scatterplot for these data using ggplot.

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

In [16]:

```r
# Set seed for reproducibility
set.seed(1)

# Generate predictor variable
x <- seq(1, 10, by = 0.5)

# Generate response variable with a non-linear relationship
y <- 2 * x^2 + 3 * x + rnorm(length(x), mean = 0, sd = 2)

# Create a data frame
data <- data.frame(x, y)

# Plot the scatterplot
ggplot(data, aes(x = x, y = y)) +
  geom_point() +
  labs(x = "Predictor", y = "Response") +
  ggtitle("Scatterplot of Non-linear Data")


# Fit linear regression model
model <- lm(y ~ x, data = data)

# Plot diagnostic plots
par(mfrow = c(2, 2))
plot(model)
```





ng departures from
n-random pattern,
indicating nonlinearity. The Cook's distance plot can be used to identify influential observations, but it may not directly indicate nonlinearity.

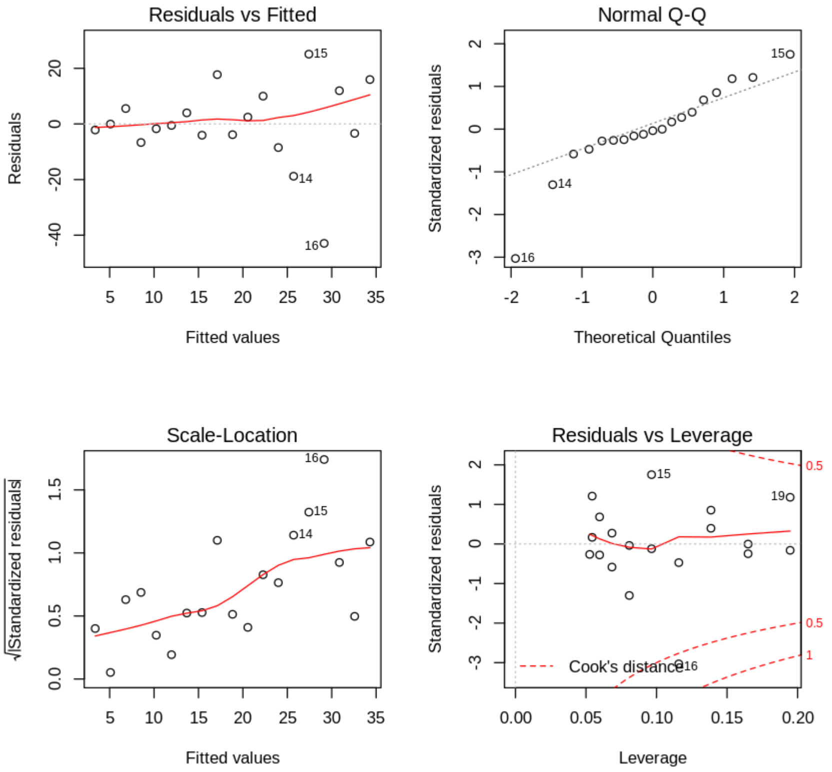
Examining these diagnostic plots, you can diagnose nonlinearity and assess the violation of the linearity assumption in the data.

### 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.

In [17]:

```
# Your Code Here
# Set seed for reproducibility
set.seed(2)

# Generate predictor variable
x <- seq(1, 10, by = 0.5)

# Generate response variable with non-constant variance
error <- abs(x) * rnorm(length(x), mean = 0, sd = 2)
y <- 3 * x + error

# Create a data frame
data <- data.frame(x, y)

# Fit linear regression model
model <- lm(y ~ x, data = data)

# Plot diagnostic plots
par(mfrow = c(2, 2))
plot(model)
```



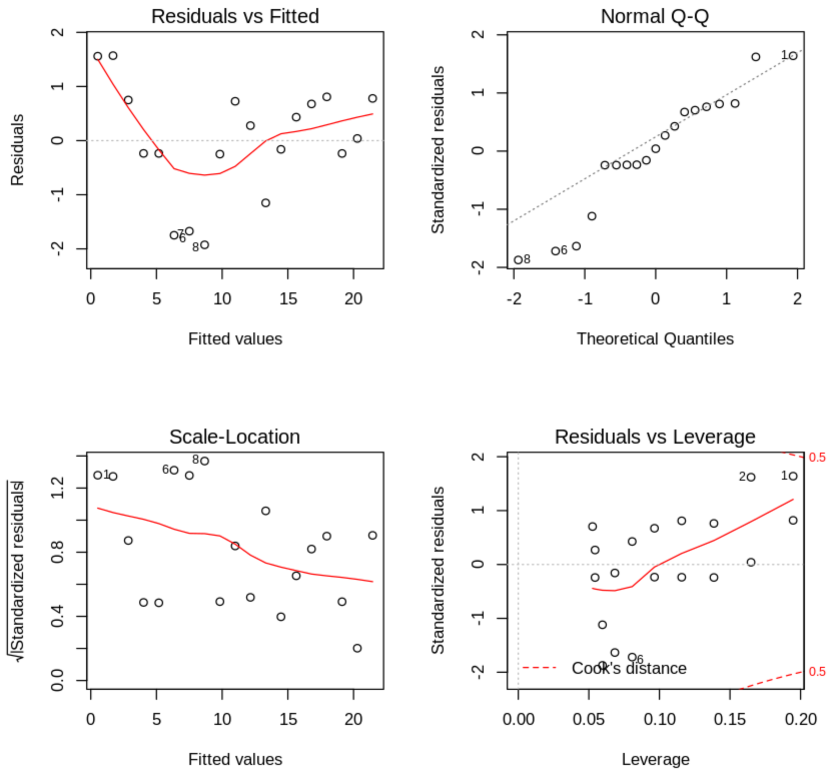
the error term is manipulated based on the absolute value of the predictor variable x. This introduces a non-constant variance, where the variability of the error term increases as the absolute value of x increases.

By plotting the diagnostic plots using the plot function with the linear model object, you can diagnose non-constant variance.

### 1. (c) Independent Errors

Repeat the above process with simulated data that violates the independent errors assumption.

In [18]:
```
# Your Code Here
# Set seed for reproducibility
set.seed(3)

# Generate predictor variable
x <- seq(1, 10, by = 0.5)

# Generate response variable with autocorrelated errors
error <- arima.sim(model = list(ar = 0.7), n = length(x))
y <- 2 * x + error

# Create a data frame
data <- data.frame(x, y)

# Fit linear regression model
model <- lm(y ~ x, data = data)

# Plot diagnostic plots
par(mfrow = c(2, 2))
plot(model)
```



The error term is generated using the arima.sim function, which creates a time series with an autoregressive (AR) structure. The autocorrelation parameter is set to 0.7, indicating a moderate positive autocorrelation.

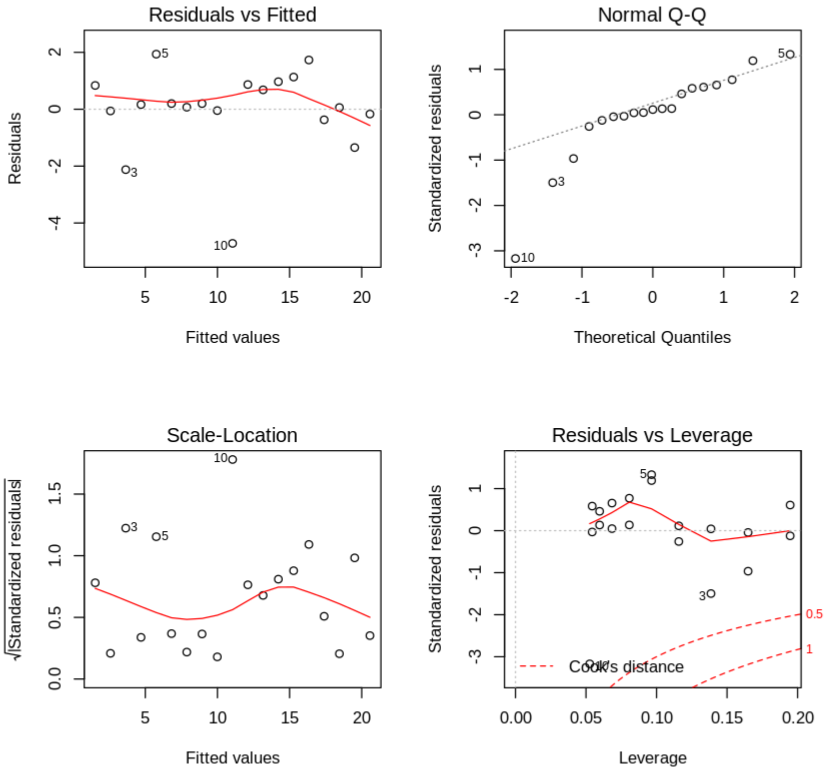
By plotting the diagnostic plots using the plot function with the linear model object, you can diagnose violations of the independent errors assumption.

### 1. (d) Normally Distributed Errors

Only one more to go! Repeat the process again but simulate the data with non-normal errors.

In [19]:

```
# Your Code Here
# Set seed for reproducibility
set.seed(4)

# Generate predictor variable
x <- seq(1, 10, by = 0.5)

# Generate response variable with non-normal errors
error <- rt(length(x), df = 3)
y <- 2 * x + error

# Create a data frame
data <- data.frame(x, y)

# Fit linear regression model
model <- lm(y ~ x, data = data)

# Plot diagnostic plots
par(mfrow = c(2, 2))
plot(model)
```



the error term is generated using the rt function, which generates random variates from a Student's t-distribution. The degrees of freedom parameter is set to 3, indicating a relatively heavy-tailed distribution compared to a normal distribution.

By plotting the diagnostic plots using the plot function with the linear model object, you can diagnose violations of the normally distributed errors assumption.

# 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" $Y$.**

$\widehat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = HY$

**2. (b) Show that $H$ is symmetric: $H = H^T$.**

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

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

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

Type *Markdown* and LaTeX: $\alpha^2$

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

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

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

Type *Markdown* and LaTeX: $\alpha^2$

**2.(e) Why is this result important in the practical use of linear regression?**

Type *Markdown* and LaTeX: $\alpha^2$

# 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.

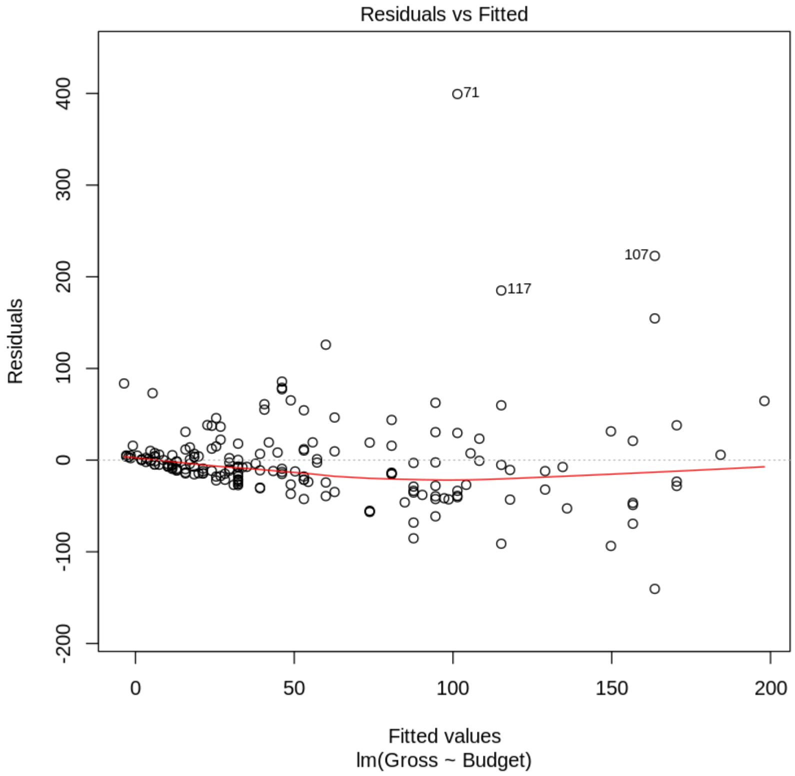
Specify whether each of the four regression model assumptions are being violated.

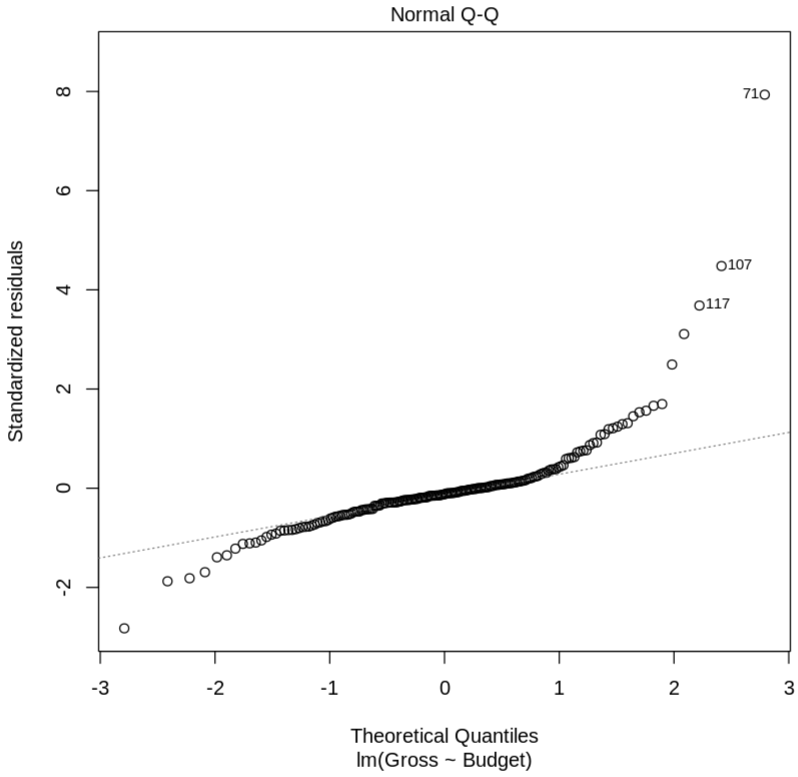
Data Source: http://www.bollymoviereviewz.com (http://www.bollymoviereviewz.com)

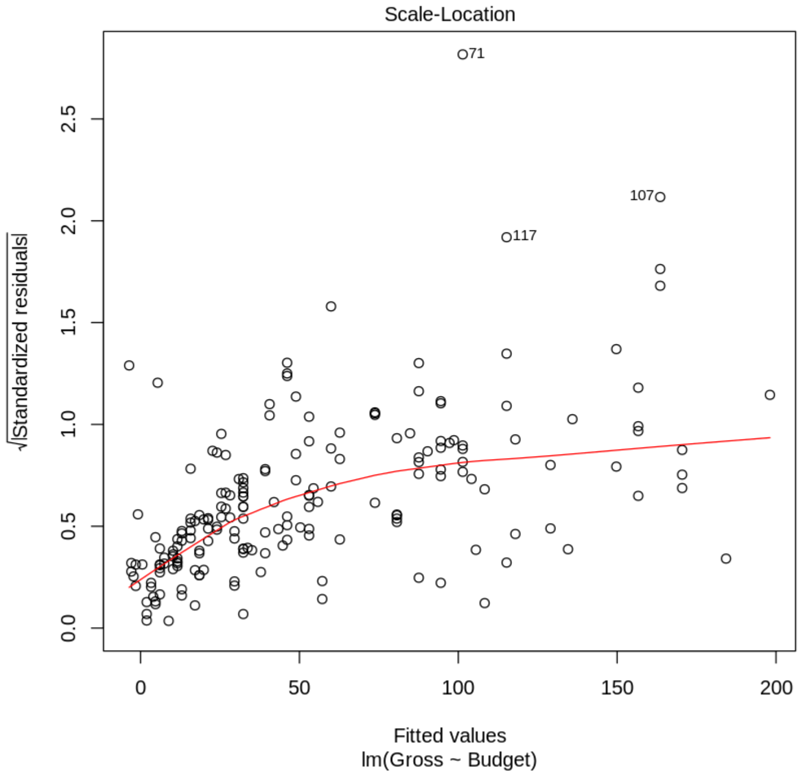
In [2]:

```
library(RCurl)
library(ggplot2)

url <- getURL("https://raw.githubusercontent.com/bzaharatos/-Statistic
bollywood <- read.csv(text = url, sep = "\t")

summary(bollywood)

lm_bollywood <- lm(Gross ~ Budget, data = bollywood)

plot(lm_bollywood)

p1 <- ggplot(bollywood, aes(x = seq_along(resid(lm_bollywood)), y = re
  geom_point() +
  stat_smooth(method = "loess", col = "#CFB87C", se = FALSE, span = 0.
  geom_hline(yintercept = 0, col = "#A2A4A3", linetype = "dashed") +
  xlab("Index") +
  ylab("Residuals") +
  ggtitle("Residual vs Index") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))

p1

n <- nrow(bollywood)
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))
```

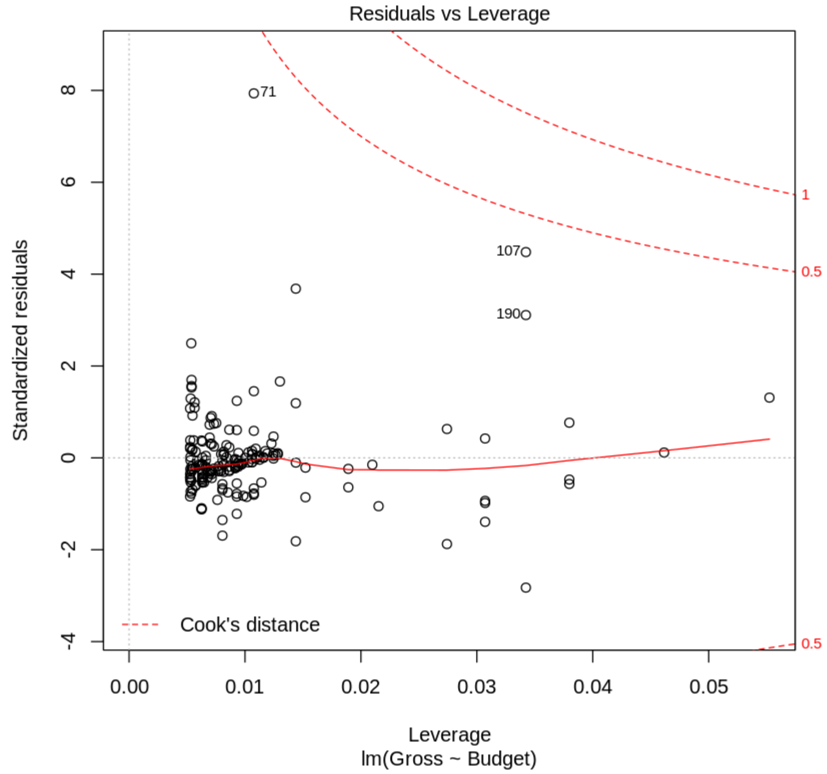
```
               Movie          Gross             Budget
 1920London       :  1   Min.   :  0.63   Min.   :  4.00
 2 States         :  1   1st Qu.:  9.25   1st Qu.: 19.00
 24(Tamil,Telugu):  1   Median : 29.38   Median : 34.50
 Aashiqui 2       :  1   Mean   : 53.39   Mean   : 45.25
 AeDilHainMushkil: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


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

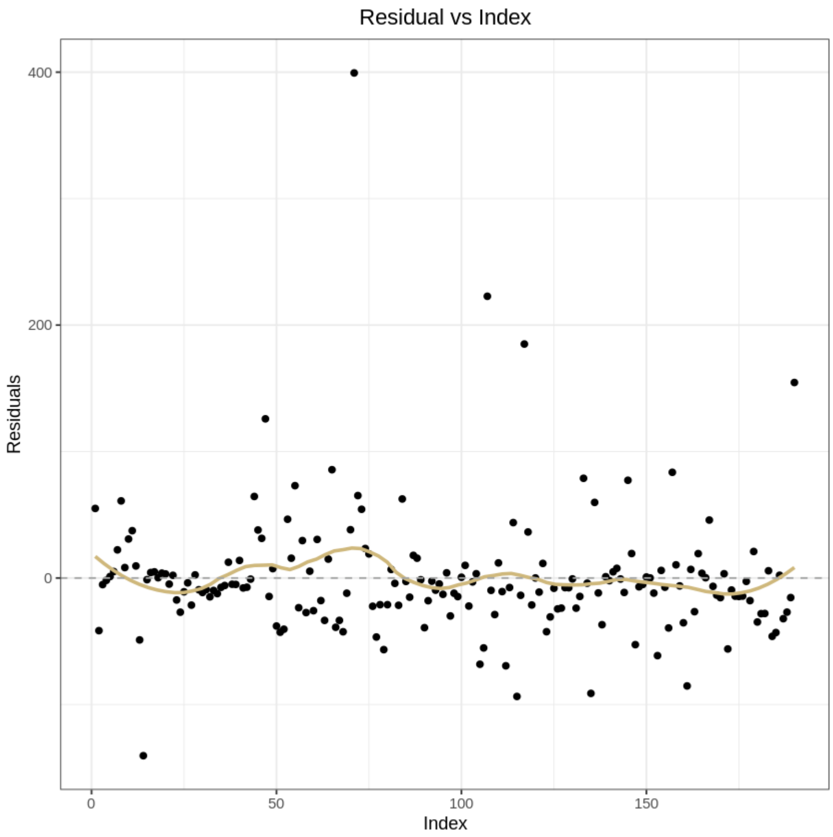

Normal Q-Q
Standardized residuals
8
6
4
2
0
-2
71
107
117
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(Gross ~ Budget)

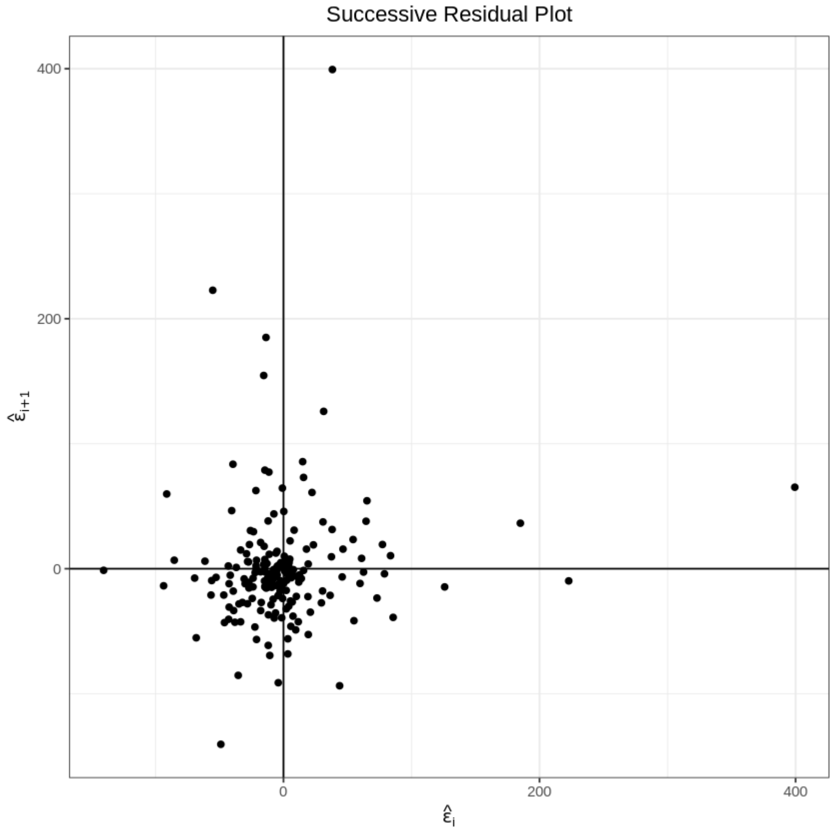


```
`geom_smooth()` using formula 'y ~ x'
```



0.111594190315052





e should be a linear
ent variable(s). The
ownward trend, indicating
nced.
sumes that the residuals
utocorrelation or
ows no discernible
or no correlation. Based
nce among successive

- Constant Assumption: There is some indication, although not very strong, of non-constant variance. This can be observed in the residual vs fitted plot, where greater variability is seen for larger fitted values compared to smaller ones.
- Normality Assumption: The normality assumption assumes that the residuals follow a normal distribution. The QQ-plot deviates from normality. However, it is likely that this

deviation is a result of the violation of linearity and constant variance assumptions.

### 3. (b) Transformations

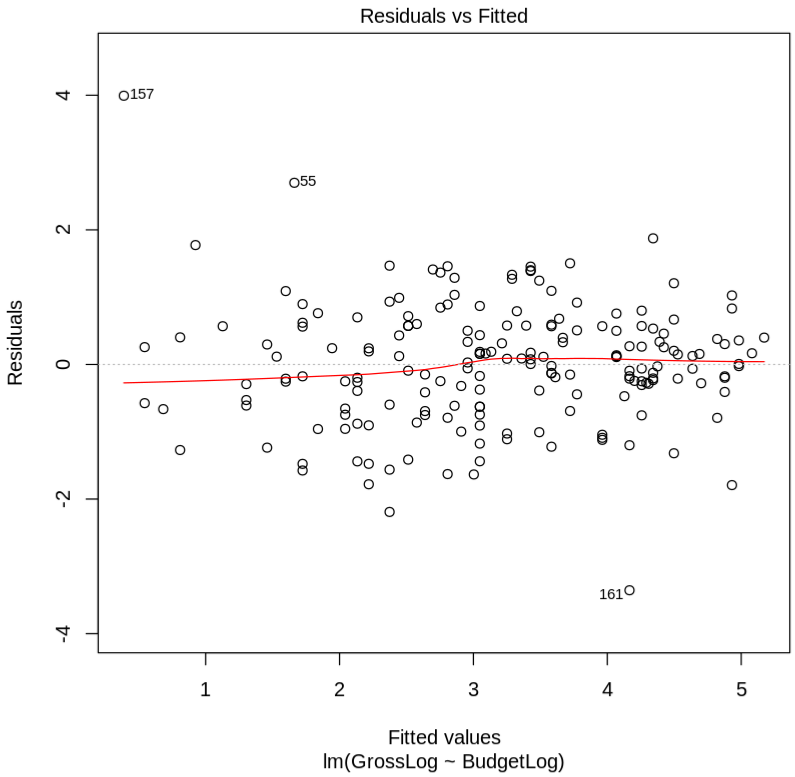
Notice that the Residuals vs. Fitted Values plot has a 'trumpet" shape to it, the points have a greater spread as the Fitted value increases. This means that there is not a constant variance, which violates the homoskedasticity assumption.

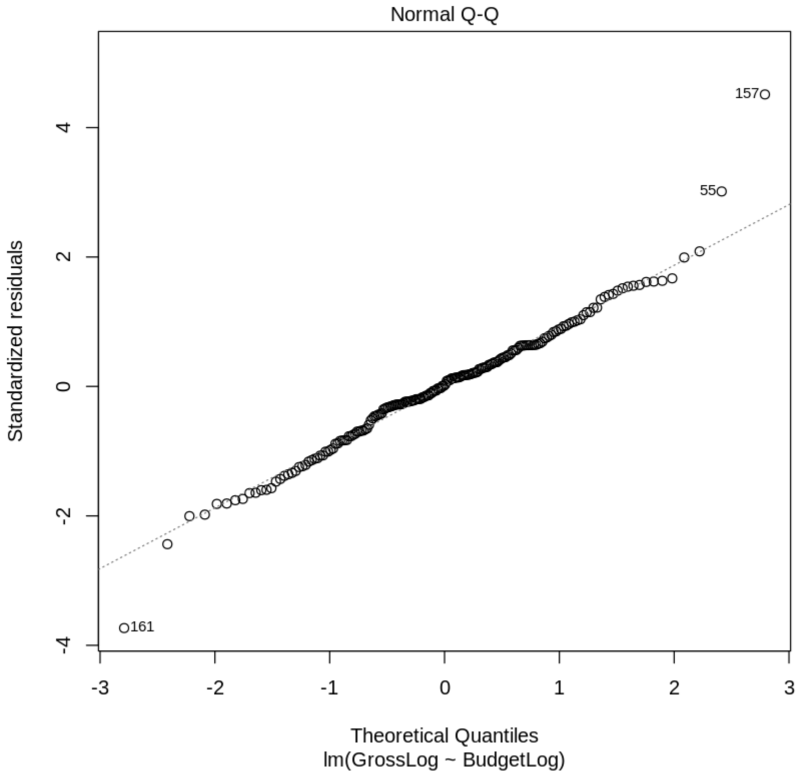
So how do we address this? Sometimes transforming the predictors or response can help stabilize the variance. Experiment with transfomrations on `Budget` and/or `Gross` so that, in the transformed scale, the relationship is approximately linear with a constant variance. Limit your transformations to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

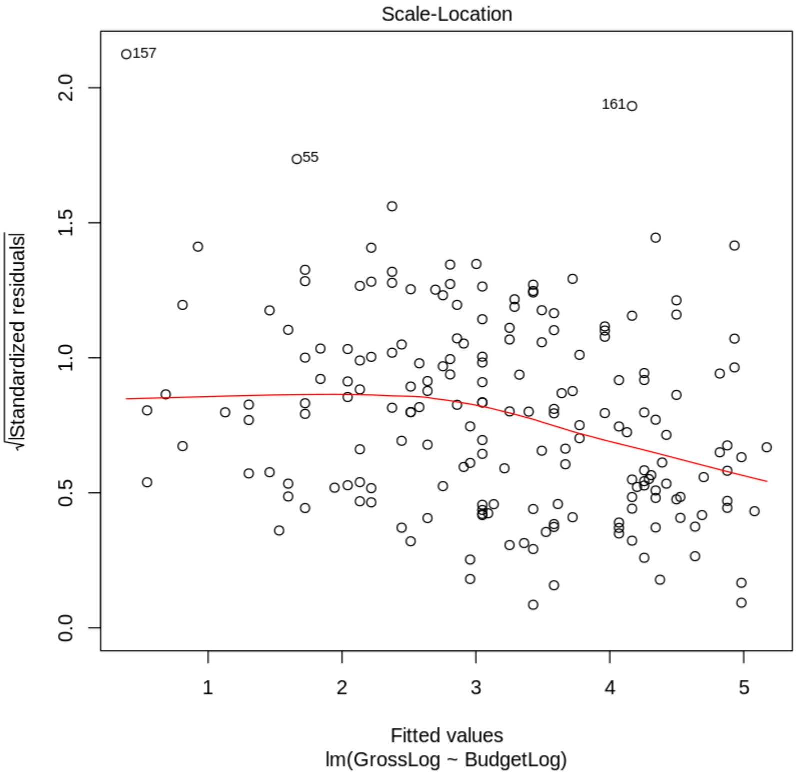
Note: There may be multiple transformations that fix this violation and give similar results. For the purposes of this problem, the transformed model doesn't have the be the "best" model, so long as it maintains both the linearity and homoskedasticity assumptions.

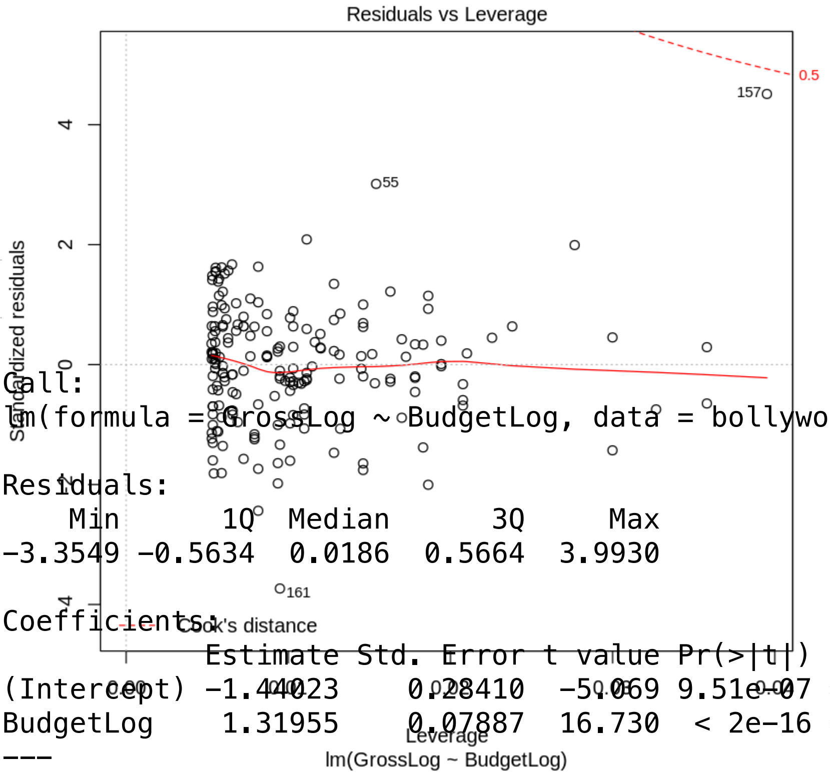
In [5]:

```
# 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)
```




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


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



e have a transformed
regression model. Write
an interpretation similar

In [4]:

```
Call:
lm(formula = GrossLog ~ BudgetLog, 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 ***
BudgetLog    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
```

4.22166668868043

Type *Markdown* and LaTeX: $\alpha^2$