Project#1

AJ Liberatore, Nathan Robinson, Reilly LeBlanc

02/21/2020

Regression

This is a group project, and students should work in a group of size 3. Include all the R code, hypothesis testing, one or two lines of explanation for any output. The report should be organized, printed, and stapled. The due date of this project is **Friday** 02/21/2020.

```
Movies = read.csv("C3 2008Movies.csv")
```

The 2008Movies file contains data on movies released in 2008.

1. Calculate a regression model to predict box office from run time. Interpret the R^2 value and test statistic for the slope in the context of this problem.

```
M1 = lm(BoxOfficeGross~RunTime, data=Movies)
summary(M1)
```

```
## 
## Call:
## lm(formula = BoxOfficeGross ~ RunTime, data = Movies)
## 
## Residuals:
##        Min         1Q     Median         3Q        Max 
## -102059739  -39266026  -20290622   17164421  457025023 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  3506911   24316122   0.144    0.885  
## RunTime       478843     226856   2.111    0.036 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 65930000 on 204 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.02137,    Adjusted R-squared:  0.01658 
## F-statistic: 4.455 on 1 and 204 DF,  p-value: 0.03601
```

Because the R^2 value of 0.02137 is low, the model is not very accurate, even though the p-value of 0.03601 suggests that the model is statistically significant.

2. Create indicator variables for the genre and MPAA rating. Use the best subsets regression to determine an appropriate regression model.

```
Genre1 = as.numeric(Movies$Genre)
MPAA = as.numeric(Movies$MPAA)

library("leaps")
Model_subset = regsubsets(BoxOfficeGross~Genre + MPAA, data=Movies)
summary(Model_subset)
```

```
## Subset selection object
## Call: regsubsets.formula(BoxOfficeGross ~ Genre + MPAA, data = Movies)
## 14 Variables  (and intercept)
##                          Forced in Forced out
## GenreAdventure               FALSE      FALSE
## GenreComedy                  FALSE      FALSE
## GenreConcert/Performance     FALSE      FALSE
## GenreDocumentary             FALSE      FALSE
## GenreDrama                   FALSE      FALSE
## GenreHorror                  FALSE      FALSE
## GenreMusical                 FALSE      FALSE
## GenreRomantic Comedy         FALSE      FALSE
## GenreThriller/Suspense       FALSE      FALSE
## GenreWestern                 FALSE      FALSE
## MPAANot Rated                FALSE      FALSE
## MPAAPG                       FALSE      FALSE
## MPAAPG-13                    FALSE      FALSE
## MPAAR                        FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          GenreAdventure GenreComedy GenreConcert/Performance GenreDocumentary
## 1  ( 1 ) "*"            " "         " "                      " "
## 2  ( 1 ) "*"            " "         " "                      " "
## 3  ( 1 ) "*"            " "         " "                      " "
## 4  ( 1 ) "*"            " "         " "                      "*"
## 5  ( 1 ) "*"            " "         " "                      "*"
## 6  ( 1 ) " "            "*"         " "                      "*"
## 7  ( 1 ) " "            "*"         "*"                      "*"
## 8  ( 1 ) " "            "*"         "*"                      "*"
##          GenreDrama GenreHorror GenreMusical GenreRomantic Comedy
## 1  ( 1 ) " "        " "         " "          " "
## 2  ( 1 ) " "        " "         " "          " "
## 3  ( 1 ) "*"        " "         " "          " "
## 4  ( 1 ) "*"        " "         " "          " "
## 5  ( 1 ) "*"        " "         "*"          " "
## 6  ( 1 ) "*"        " "         " "          "*"
## 7  ( 1 ) "*"        " "         " "          "*"
## 8  ( 1 ) "*"        "*"         " "          "*"
##          GenreThriller/Suspense GenreWestern MPAANot Rated MPAAPG MPAAPG-13
## 1  ( 1 ) " "                    " "          " "           " "    " "
## 2  ( 1 ) " "                    " "          " "           " "    " "
## 3  ( 1 ) " "                    " "          " "           " "    " "
## 4  ( 1 ) " "                    " "          " "           " "    " "
## 5  ( 1 ) " "                    " "          " "           " "    " "
## 6  ( 1 ) "*"                    " "          " "           " "    " "
## 7  ( 1 ) "*"                    " "          " "           " "    " "
## 8  ( 1 ) "*"                    " "          " "           " "    " "
##          MPAAR
## 1  ( 1 ) " "
## 2  ( 1 ) "*"
## 3  ( 1 ) "*"
## 4  ( 1 ) "*"
## 5  ( 1 ) "*"
## 6  ( 1 ) "*"
## 7  ( 1 ) "*"
## 8  ( 1 ) "*"
```

```r
Comedy = as.numeric(Movies$Genre == "Comedy")
Concert = as.numeric(Movies$Genre == "Concert/Performance")
Documentary = as.numeric(Movies$Genre == "Documentary")
Drama = as.numeric(Movies$Genre == "Drama")
Horror = as.numeric(Movies$Genre == "Horror")
RomCom = as.numeric(Movies$Genre == "Romantic Comedy")
Thriller = as.numeric(Movies$Genre == "Thriller/Suspense")
MPAAR = as.numeric(Movies$MPAA == "R")

Model_full= lm(BoxOfficeGross~Comedy + Documentary + Drama + Horror + RomCom + Thriller + MPAAR,
data=Movies)
summary(Model_full)
```

```
##
## Call:
## lm(formula = BoxOfficeGross ~ Comedy + Documentary + Drama +
##     Horror + RomCom + Thriller + MPAAR, data = Movies)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -87403806 -34956378 -15322486  15599782 441377138
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  91938923    9290754   9.896  < 2e-16 ***
## Comedy      -32622258   12217608  -2.670 0.008213 **
## Documentary -63151227   23048950  -2.740 0.006708 **
## Drama       -50116653   13090337  -3.829 0.000173 ***
## Horror      -25167335   23297041  -1.080 0.281331
## RomCom      -40335879   22053247  -1.829 0.068901 .
## Thriller    -40776589   17191709  -2.372 0.018657 *
## MPAAR       -20904677    9660465  -2.164 0.031666 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63550000 on 198 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.1174, Adjusted R-squared:  0.08621
## F-statistic: 3.763 on 7 and 198 DF,  p-value: 0.0007475
```

```
Model_best = lm(BoxOfficeGross~Comedy + Documentary + Drama + Thriller + MPAAR, data=Movies)
summary(Model_best)
```

```
##
## Call:
## lm(formula = BoxOfficeGross ~ Comedy + Documentary + Drama +
##     Thriller + MPAAR, data = Movies)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -78418169 -37089654 -15088470  15318455 450362775
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  82953286    8108581  10.230  < 2e-16 ***
## Comedy      -23261221   11247471  -2.068 0.039914 *
## Documentary -53919296   22656062  -2.380 0.018257 *
## Drama       -40706552   12182281  -3.341 0.000994 ***
## Thriller    -31181374   16499744  -1.890 0.060230 .
## MPAAR       -22013000    9601360  -2.293 0.022906 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63870000 on 200 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.09966,    Adjusted R-squared:  0.07715
## F-statistic: 4.428 on 5 and 200 DF,  p-value: 0.000758
```

```
genreslist = cbind(cbind(cbind(Comedy, Documentary), Drama), Thriller)
Movies = cbind(cbind(Movies, genreslist), MPAAR)
```

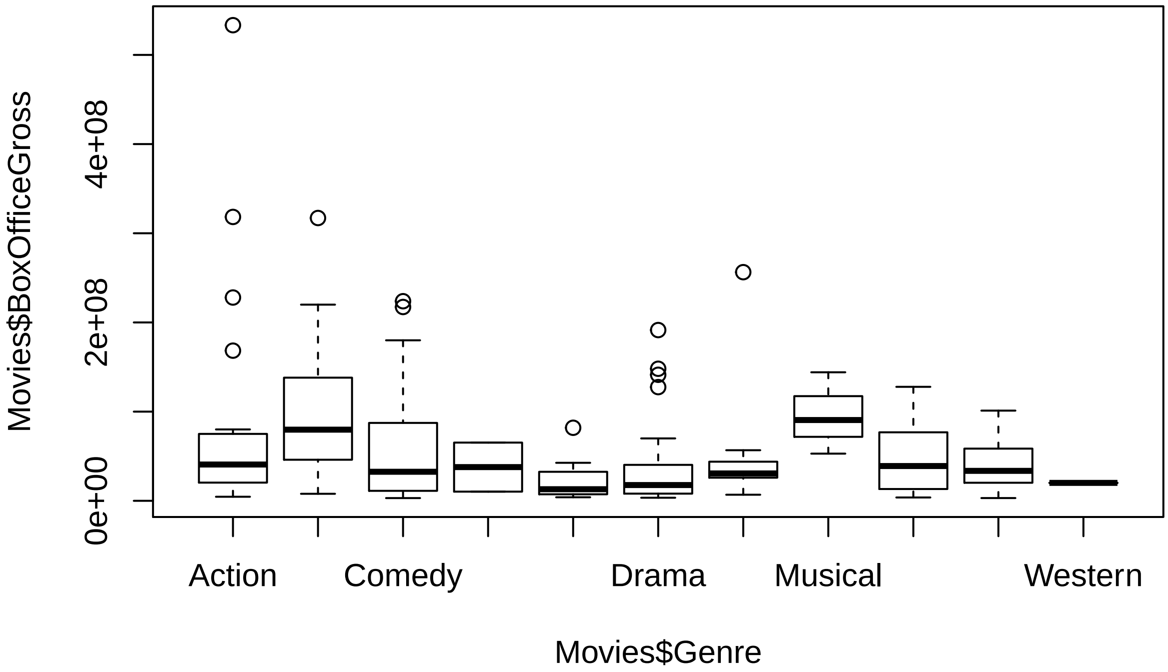
a. Validate the model assumptions.

Linear Relationship

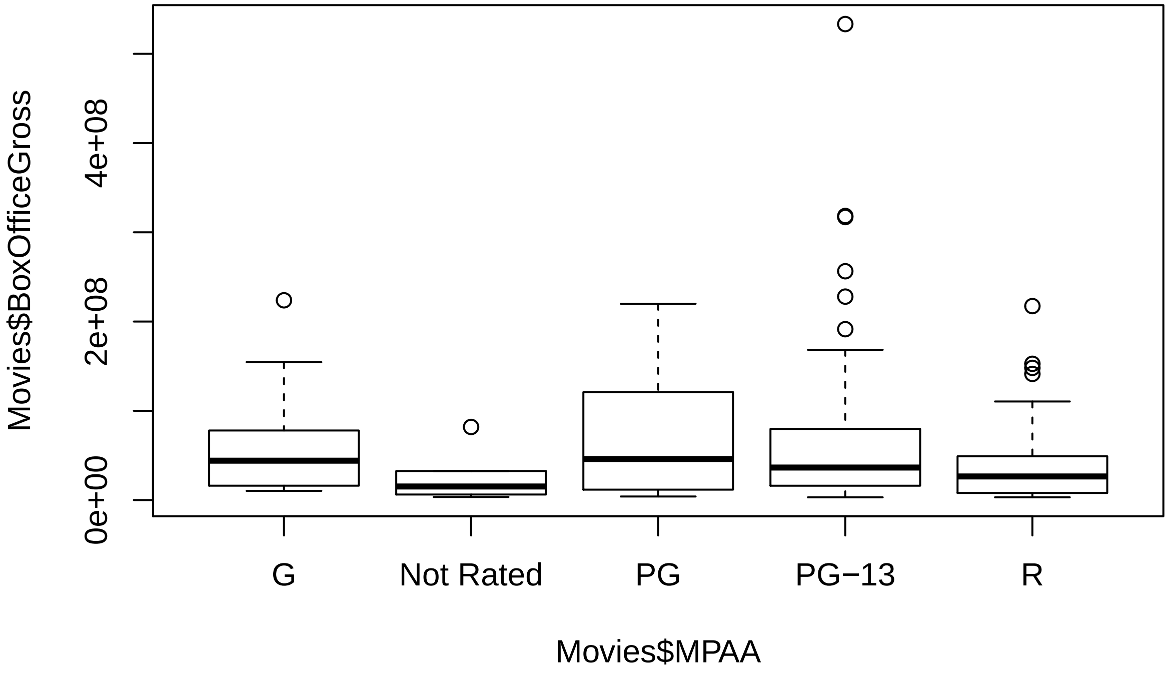
By examining the coefficient of determination, it can be seen that there is no linear relationship.

Multivariate Normality.

```
boxplot(Movies$BoxOfficeGross~Movies$Genre)
```



```
boxplot(Movies$BoxOfficeGross~Movies$MPAA)
```



By examining the boxplots, it can be seen that there is not multivariate normality.

Little or no Multicolinearity

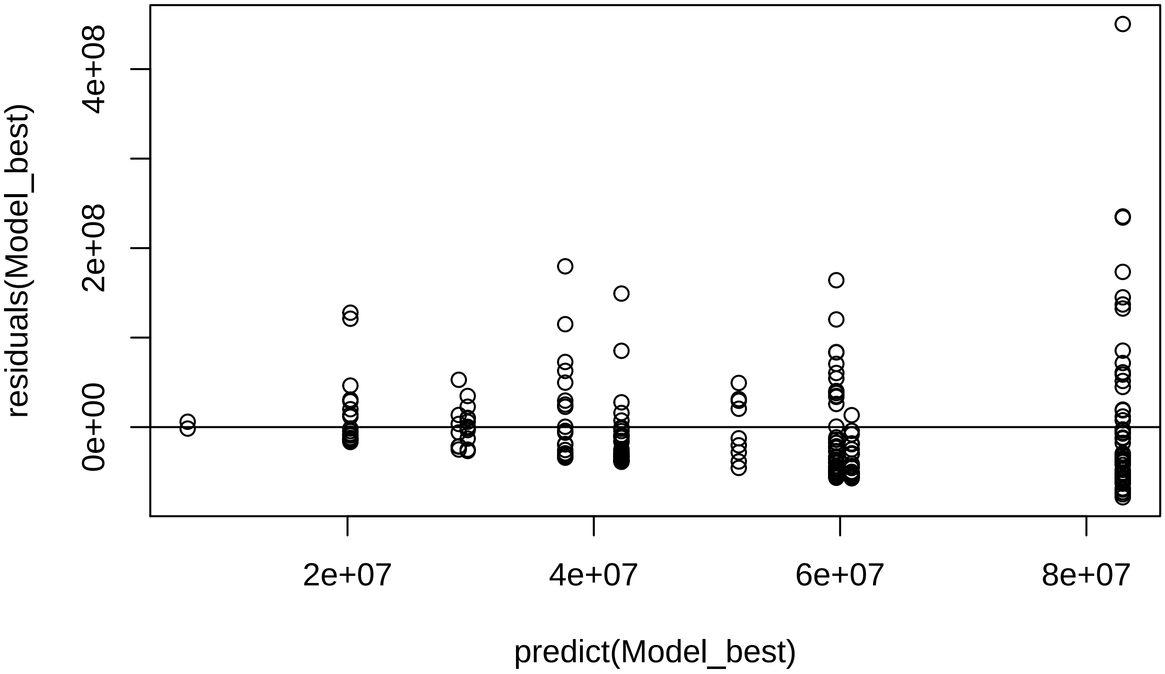
```
library(mctest)
imcdata = data.matrix(Movies[19:23])
imcdiag(x = imcdata, y = Movies$BoxOfficeGross)
```

```
##
## Call:
## imcdiag(x = imcdata, y = Movies$BoxOfficeGross)
##
##
## All Individual Multicollinearity Diagnostics Result
##
##                 VIF    TOL      Wi      Fi Leamer   CVIF Klein   IND1   IND2
## Comedy      1.3441 0.7440 17.2904 23.1685 0.8626 1.3089     1 0.0148 1.6390
## Documentary 1.0830 0.9233  4.1723  5.5908 0.9609 1.0547     0 0.0184 0.4908
## Drama       1.3198 0.7577 16.0711 21.5347 0.8704 1.2852     1 0.0151 1.5514
## Thriller    1.2052 0.8297 10.3115 13.8171 0.9109 1.1736     1 0.0165 1.0901
## MPAAR       1.0370 0.9643  1.8616  2.4945 0.9820 1.0099     0 0.0192 0.2287
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## Thriller , coefficient(s) are non-significant may be due to multicollinearity
##
## R-square of y on all x: 0.0997
##
## * use method argument to check which regressors may be the reason of collinearity
## ===================================
```

Given the F-G test, GenreComedy, GenreDrama, and GenreThriller/Suspense are all sources of multicollinearity.

Variance of error terms are similar

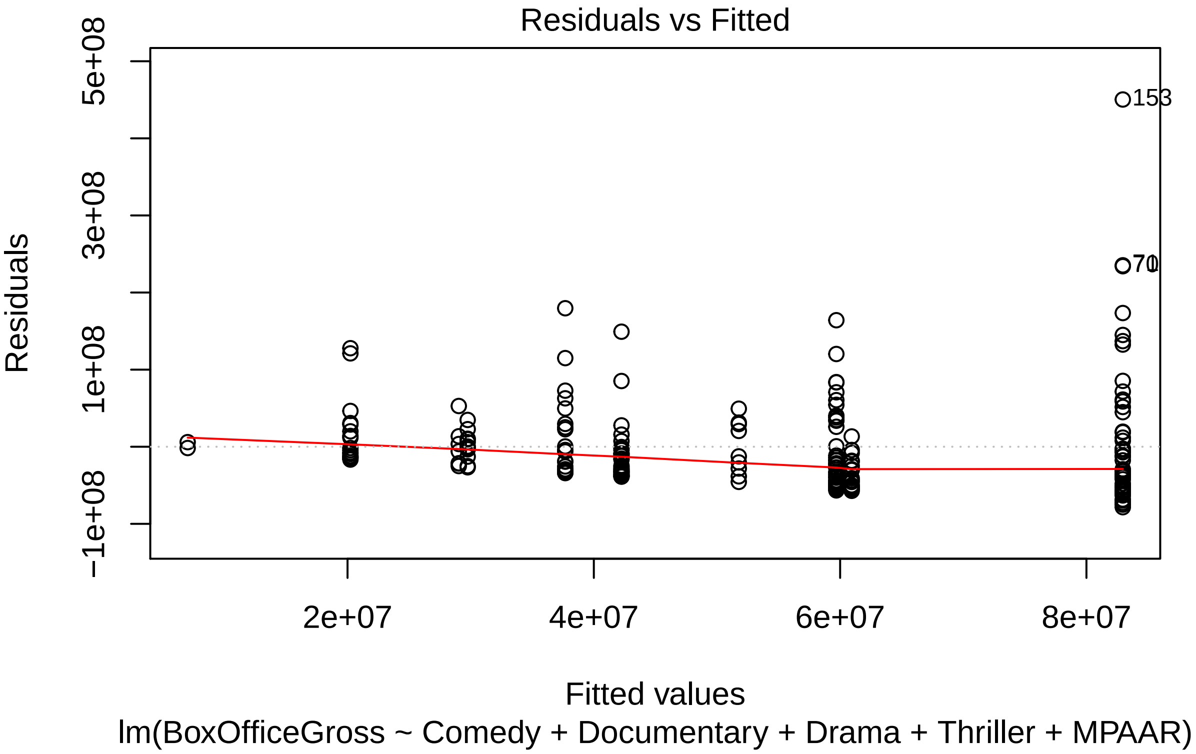
```r
plot(residuals(Model_best)~predict(Model_best))
abline(lm(residuals(Model_best)~predict(Model_best)))
```

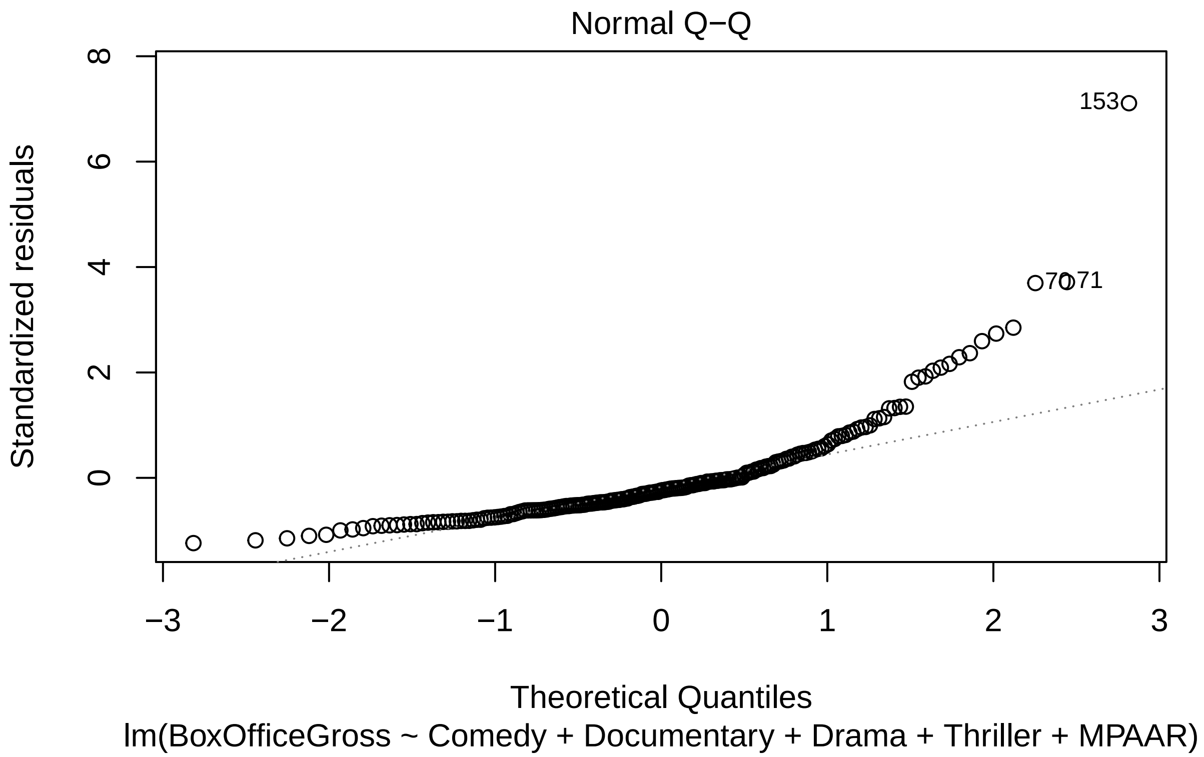


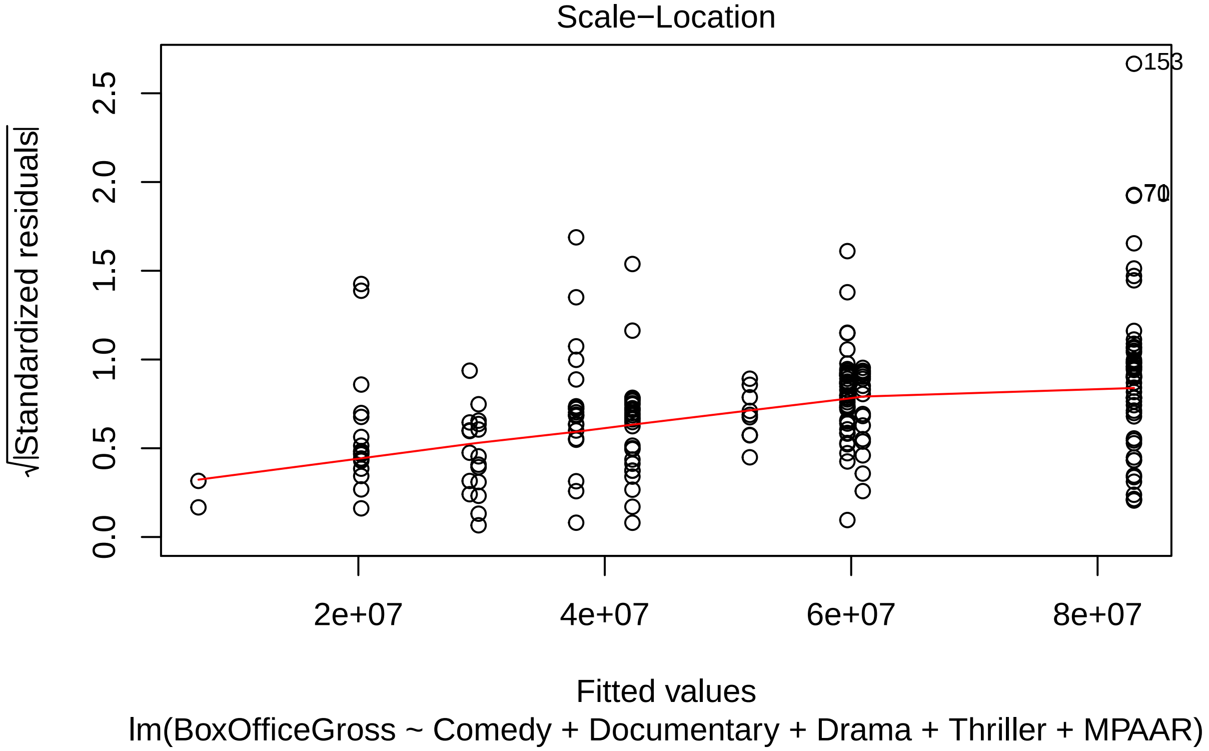
Given a plot of residuals vs predicted values, it appears that the error terms are similar and variance is not affected by how large a predicted value is.

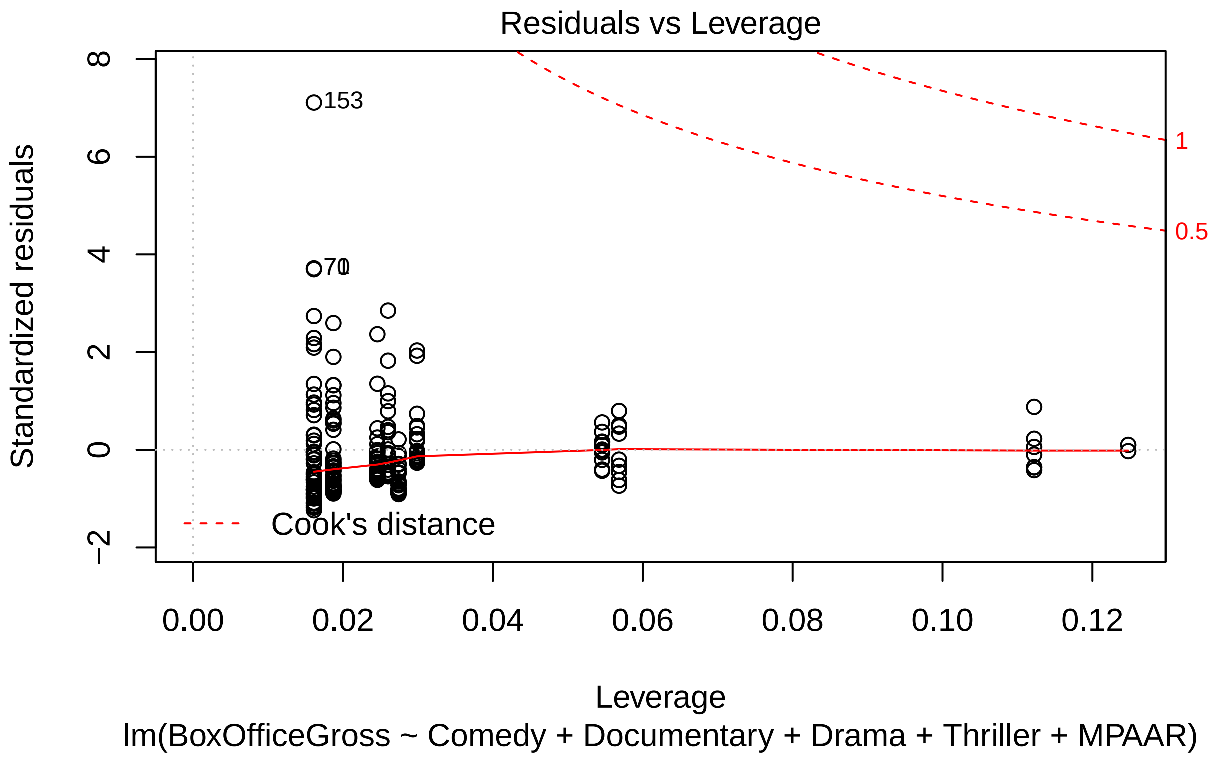
b. Look at residual plots and check for heteroskedasticity (unequal variance), multicollinearity, correlation of errors, and outliers. Transform the data if it is appropriate.

```
plot(Model_best)
```




Normal Q−Q
Standardized residuals
0
2
4
6
8
153
70
71
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(BoxOfficeGross ~ Comedy + Documentary + Drama + Thriller + MPAAR)

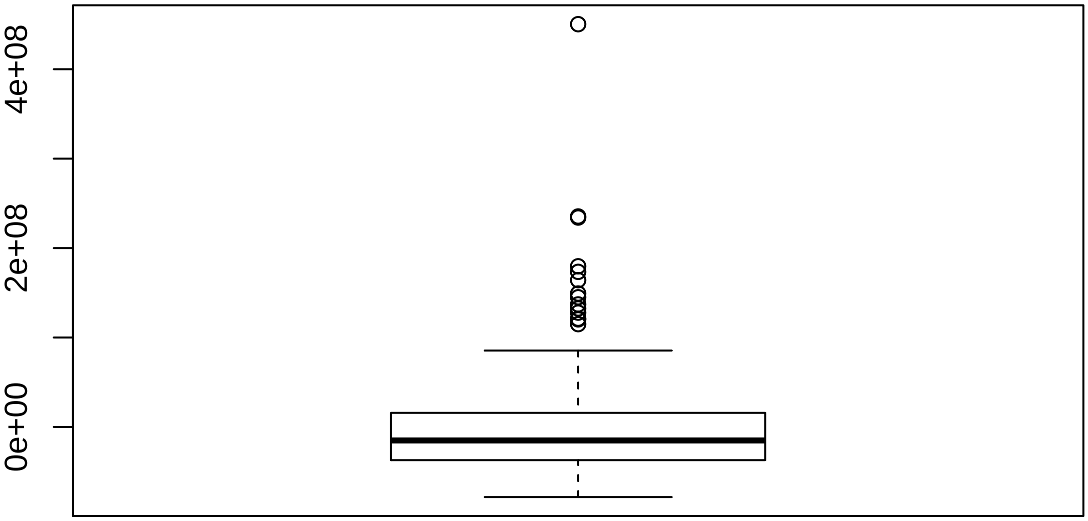

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
153
70
2e+07
4e+07
6e+07
8e+07
Fitted values
lm(BoxOfficeGross ~ Comedy + Documentary + Drama + Thriller + MPAAR)



```r
summary(lm(residuals(Model_best)~predict(Model_best)))
```

```
##
## Call:
## lm(formula = residuals(Model_best) ~ predict(Model_best))
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -78418169 -37089654 -15088470  15318455 450362775
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -2.617e-09  1.217e+07       0        1
## predict(Model_best)  4.619e-17  2.104e-01       0        1
##
## Residual standard error: 63240000 on 204 degrees of freedom
## Multiple R-squared:  1.694e-33,  Adjusted R-squared:  -0.004902
## F-statistic: 3.455e-31 on 1 and 204 DF,  p-value: 1
```

```r
boxplot(residuals(Model_best))
```



Heteroskedasticity: It was found in part a that the variance is equal, and the Residual vs. Fitted plot further shows that this is true.

Multicollinearity: It was found in part a that there are some terms which cause multicollinearity, further shown in the residual plots.

Correlation of Errors: It was found that there is no correlation for errors, given the summary statistics for the linear model for residuals vs. predicted values.

Outliers: Given the boxplot, it appears that there are quite a few outliers in the data in terms of residuals.

Transformation: Given that the independent variables in this case were all 0 or 1, it was found that most transformations would yield little to no result. However, a log transform of the response variable will result in a slightly better model. The new model is shown below:

```
Model_best2 = lm(log(BoxOfficeGross)~Comedy  + Documentary + Drama + Thriller + MPAAR, data=Movies)
summary(Model_best2)
```

```
##
## Call:
## lm(formula = log(BoxOfficeGross) ~ Comedy + Documentary + Drama +
##     Thriller + MPAAR, data = Movies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.41019 -0.81542  0.08627  0.79797  2.35707
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.7376     0.1381 128.404  < 2e-16 ***
## Comedy       -0.4223     0.1916  -2.204 0.028686 *
## Documentary  -1.1104     0.3860  -2.877 0.004450 **
## Drama        -0.8001     0.2075  -3.855 0.000156 ***
## Thriller     -0.3282     0.2811  -1.168 0.244325
## MPAAR        -0.4452     0.1636  -2.722 0.007063 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.088 on 200 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.1259, Adjusted R-squared:  0.1041
## F-statistic: 5.762 on 5 and 200 DF,  p-value: 5.42e-05
```

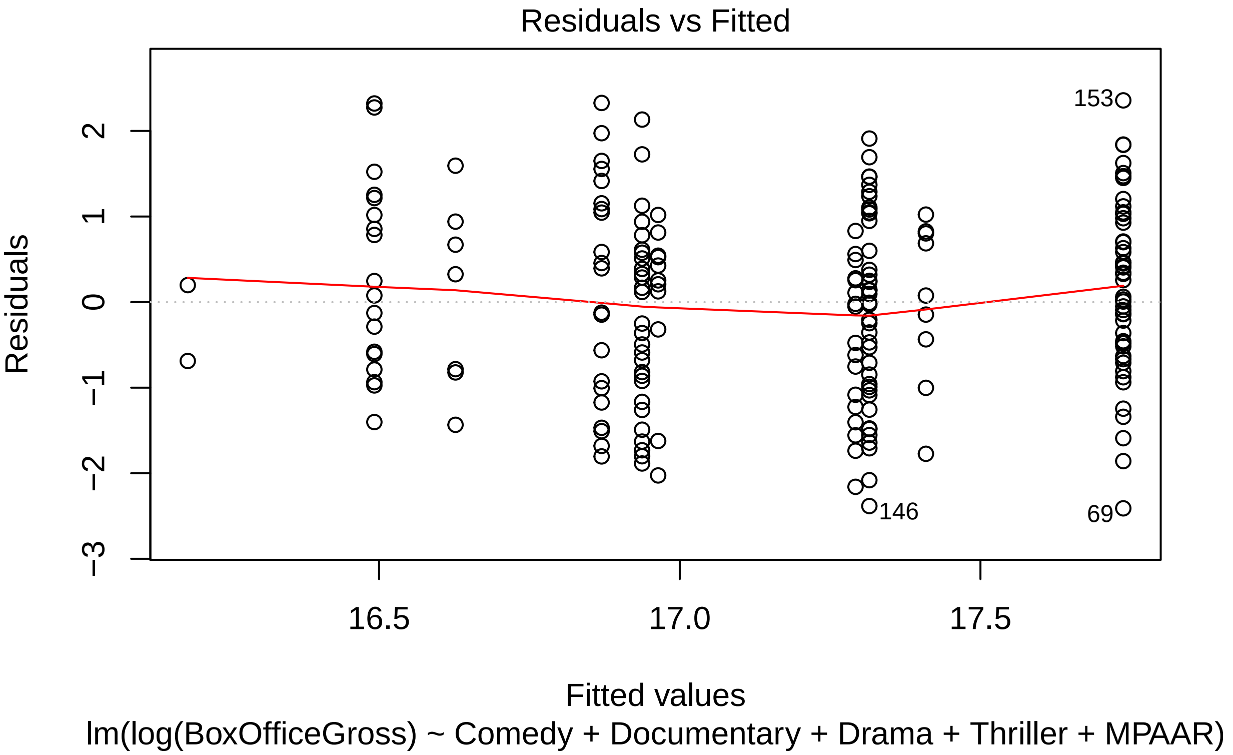
c. Submit your suggested least squares regression formula along with a limited number of appropriate graphs that provide justification for your model. Describe why you believe this model is the best.

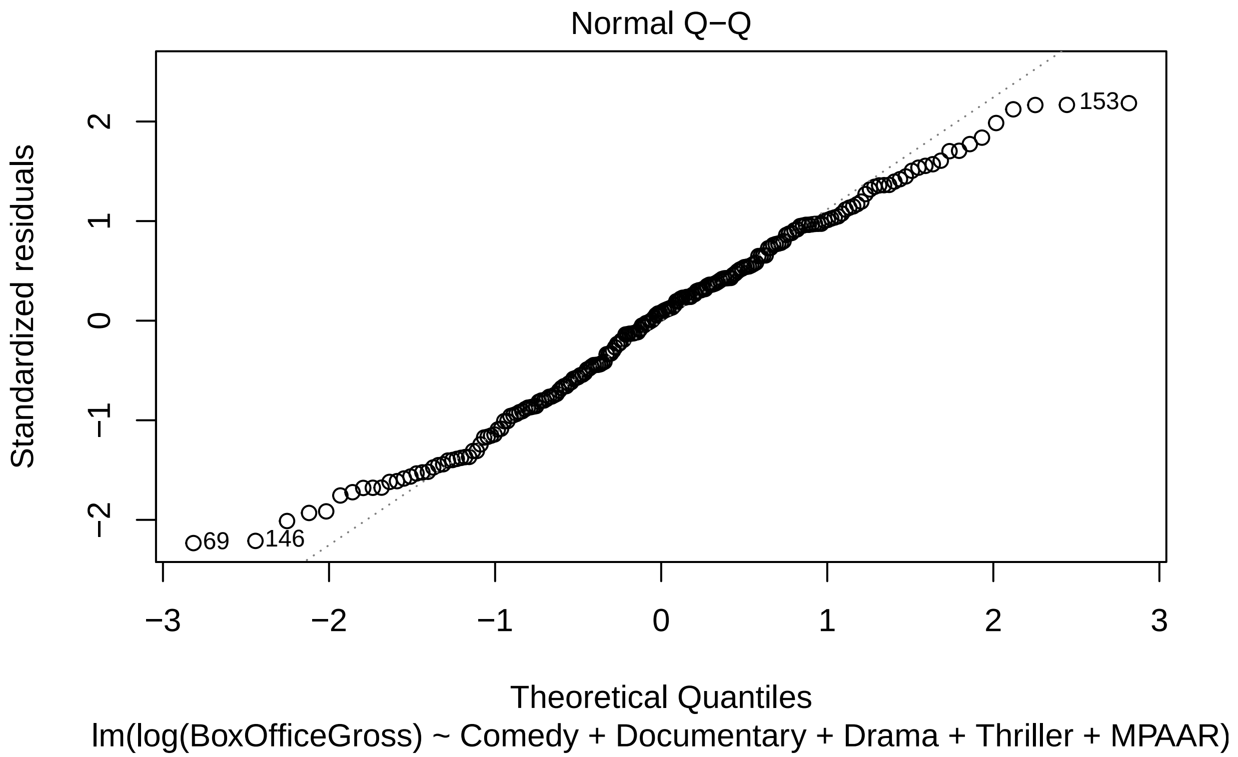
The best least squares regression formula is as follows:

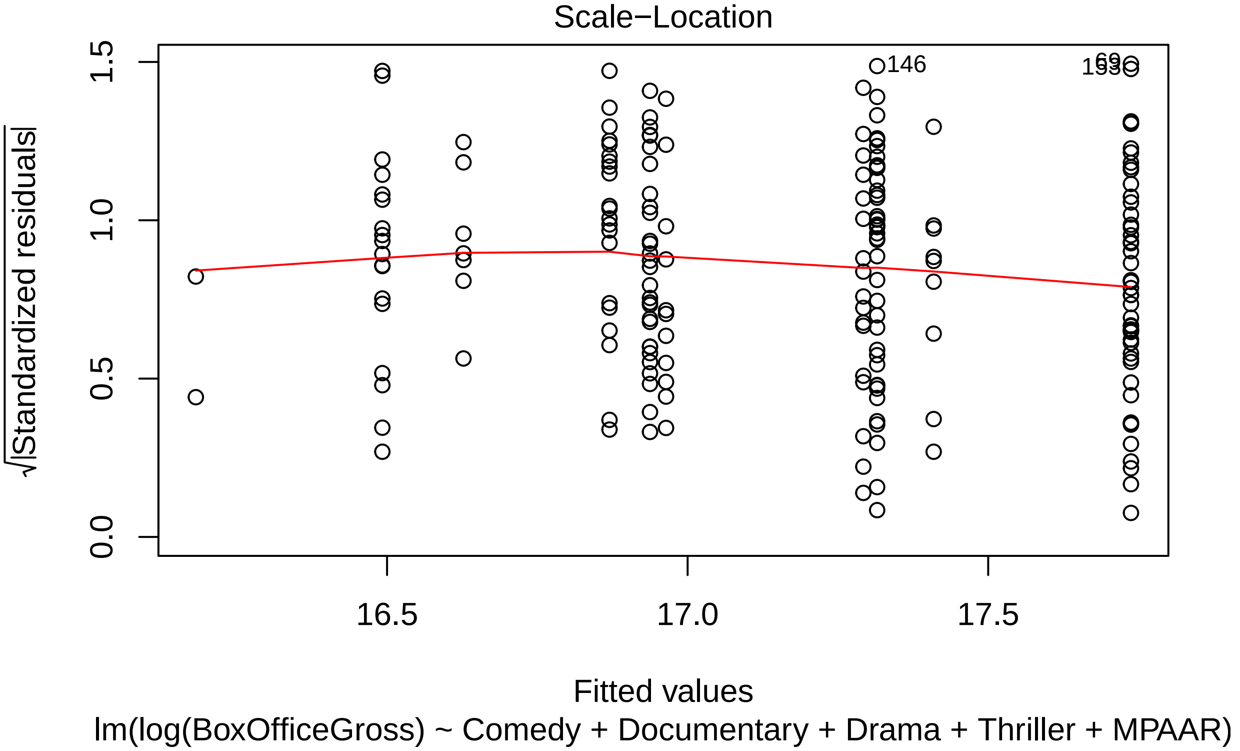
$$log(BoxOfficeGross) = 17.738 - 0.422(Comedy) - 1.110(Documentary) - 0.800(Drama) - 0.328(Thriller) - 0.445(MPAAR)$$

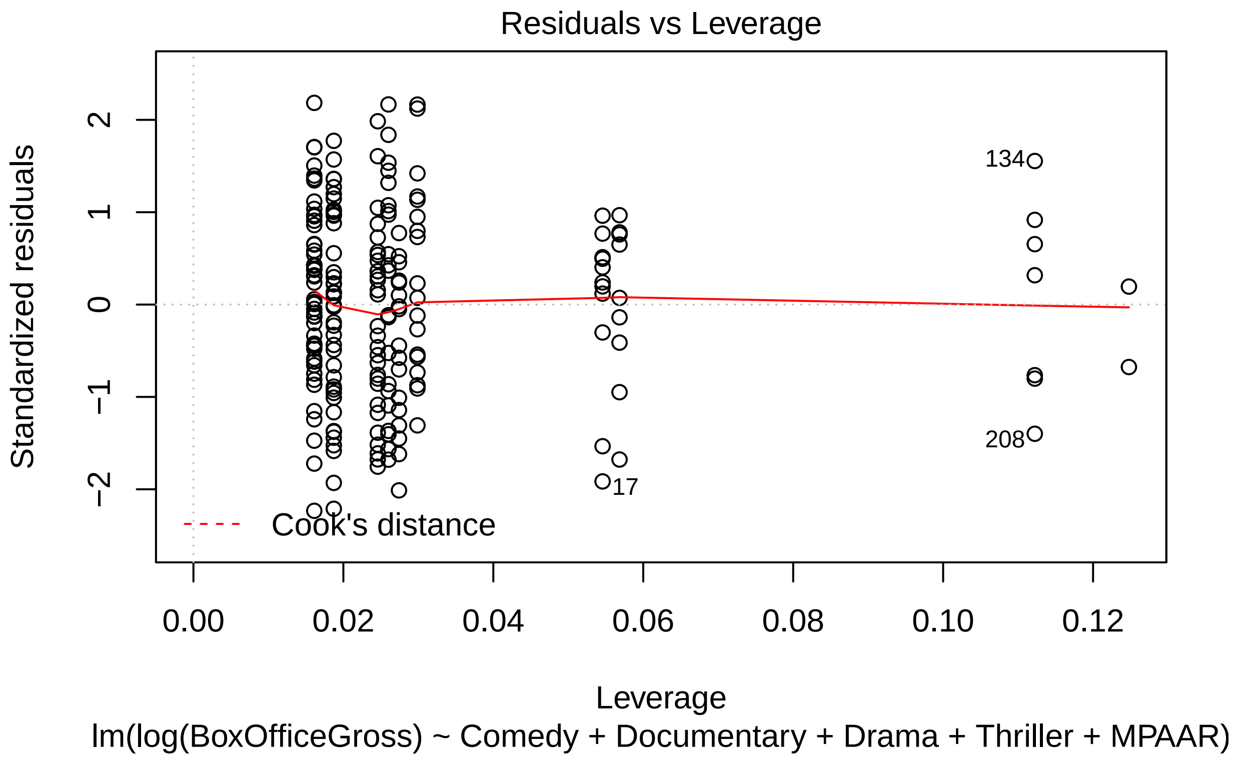
This model is better than the previous model, given a higher coefficient of determination and more normal of a residual distribution.

```
plot(Model_best2)
```

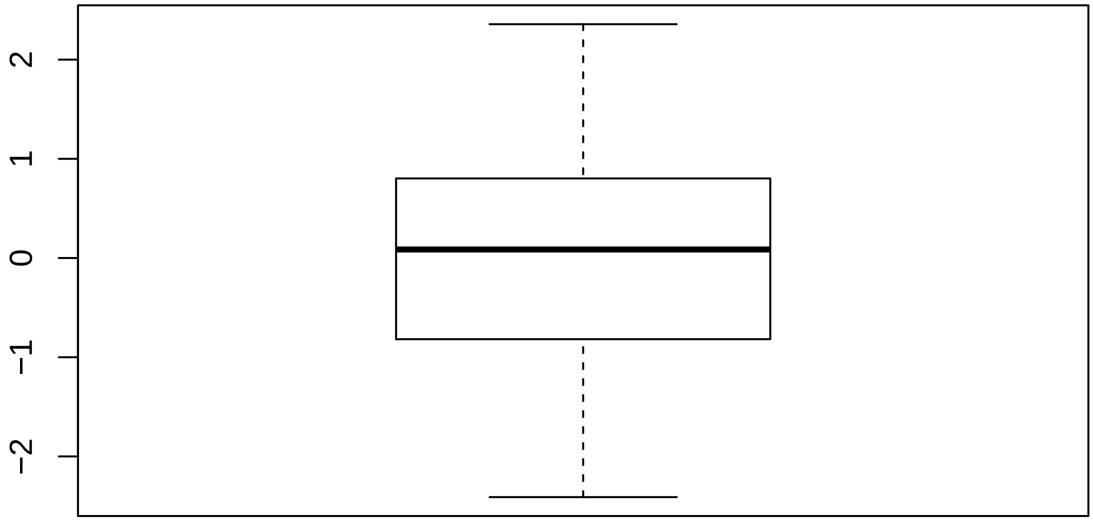



Normal Q−Q
Standardized residuals
2
1
0
−1
−2
153
69
146
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(log(BoxOfficeGross) ~ Comedy + Documentary + Drama + Thriller + MPAAR)


Scale−Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
146
69
153
16.5
17.0
17.5
Fitted values
lm(log(BoxOfficeGross) ~ Comedy + Documentary + Drama + Thriller + MPAAR)



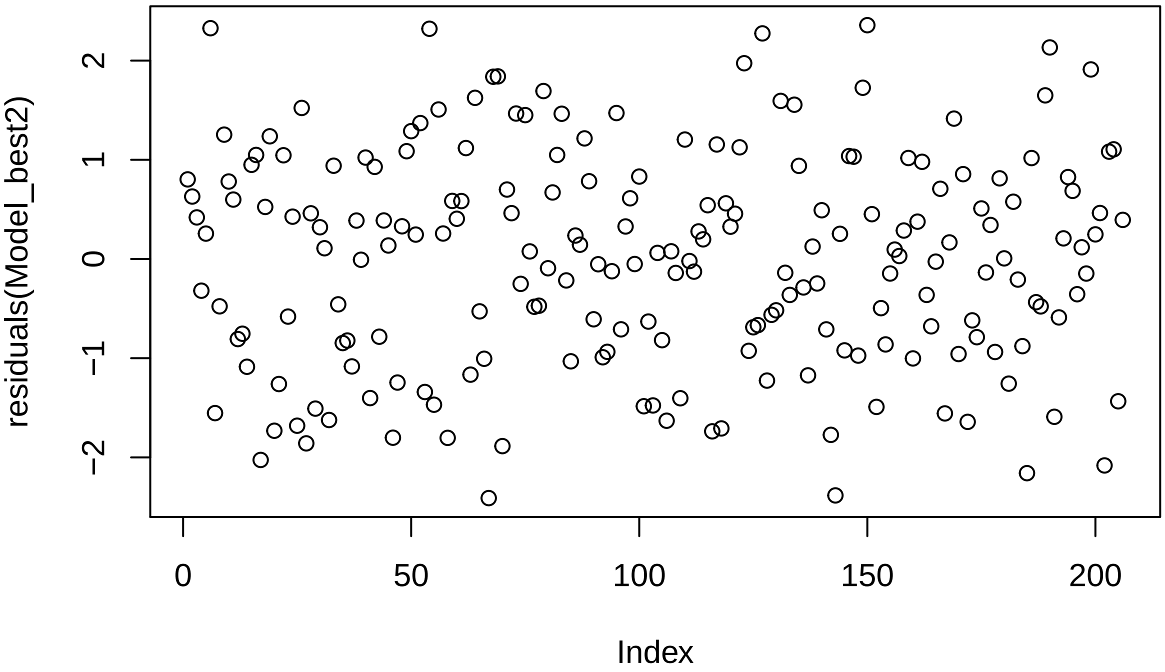
```r
boxplot(residuals(Model_best2))
```



The qq-plot specifically shows normal residual distribution, and the variance of the residuals does not increase given an increase in the predicted value. A boxplot further proves this, given the shape of the boxplot and the fact that there are no outliers.

d. Test the overall model adequacy.

```r
plot(residuals(Model_best2))
```



Overall, this model is more adequate than the previous one, but that doesn't say much given a very small R^2 value, which was 0.1259. Though residuals are now normally distributed, and the residual vs. order plot looks completely randomized, not much of a difference is made in terms of model adequacy.

3. Conduct an extra sum of squares test to determine if one or more interaction terms (or quadratic terms) should be included in the model. You can choose any other terms to test.
   $H_0 : \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$
   H_a : At least one of the β_i is not zero.

```
Model_interaction = lm(log(BoxOfficeGross)~Comedy  + Documentary + Drama + Thriller + MPAAR + Comedy*MP
anova(Model_best2, Model_interaction)
```

```
## Analysis of Variance Table
##
## Model 1: log(BoxOfficeGross) ~ Comedy + Documentary + Drama + Thriller +
##     MPAAR
## Model 2: log(BoxOfficeGross) ~ Comedy + Documentary + Drama + Thriller +
##     MPAAR + Comedy * MPAAR + Documentary * MPAAR + Drama * MPAAR +
##     Thriller * MPAAR
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    200 236.78
## 2    196 227.02  4    9.7599 2.1066 0.08144 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Given the large p-value, at the .95 level of confidence the null hypothesis is not rejected; thus, there is enough statistical evidence to conclude that interaction terms make no difference to the linear model.

It is of note that these interaction terms were used because there can only be interaction between genre and rating; a film can only be one genre and one rating, not several.

4. Test whether average run time is the same for different Genre. Clearly show your hypothesis test.

   Using $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7 = \mu_8 = \mu_9 = \mu_{10}$ and $H_a : \mu_i$ are not all equal.

```
summary(aov(RunTime~Genre, data = Movies))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## Genre        10  26267  2626.7   8.871 5.51e-12 ***
## Residuals   197  58333   296.1                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 1 observation deleted due to missingness
```

With a p-value of 5.51×10^{-12} we have sufficient evidence to reject the null hypothesis and conclude that the average runtime for differnet movie genres is not the same.

5. Check equality of variance of run time for Genre type.

   Using Barlett's test with $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2 = \sigma_6^2 = \sigma_7^2 = \sigma_8^2 = \sigma_9^2$ and $H_a : \sigma_i^2$ are not all equal.

```
Movies2 = Movies[-8,] # We omit western because it is the only oberservation for that genre.
bartlett.test(RunTime~Genre, data = Movies2)
```

```
## 
##  Bartlett test of homogeneity of variances
## 
## data:  RunTime by Genre
## Bartlett's K-squared = 47.651, df = 9, p-value = 2.967e-07
```

With a p-value of 2.967×10^{-7} we have suffecint evidence to reject the null hypothesis and conclude that the variance in runtime for differnet movie genres is not the same.