# Regression Modeling with Movie Dataset

Gautham Meenakshisundaram

1/15/2020

## PART 1: DATA

The dataset contains basic information, reviews, ratings and acknowledgements of movies collected through random sampling from Rotten Tomatoes and IMDB.

In a quantitative study like this, generalizability applies and the results from this sample can be extended further to population as the data is collected through random sampling. Moreover, random sampling also helps to determine the accurate degree of casuality.

```
load('movies.Rdata')
head(movies)
```

```
##                   title   title_type       genre runtime mpaa_rating
## 1          Filly Brown Feature Film       Drama      80           R
## 2             The Dish Feature Film       Drama     101       PG-13
## 3  Waiting for Guffman Feature Film      Comedy      84           R
## 4 The Age of Innocence Feature Film       Drama     139          PG
## 5          Malevolence Feature Film      Horror      90           R
## 6          Old Partner  Documentary Documentary      78     Unrated
##                     studio thtr_rel_year thtr_rel_month thtr_rel_day
## 1      Indomina Media Inc.          2013              4           19
## 2    Warner Bros. Pictures          2001              3           14
## 3   Sony Pictures Classics          1996              8           21
## 4        Columbia Pictures          1993             10            1
## 5 Anchor Bay Entertainment          2004              9           10
## 6       Shcalo Media Group          2009              1           15
##   dvd_rel_year dvd_rel_month dvd_rel_day imdb_rating imdb_num_votes
## 1         2013             7          30         5.5            899
## 2         2001             8          28         7.3          12285
## 3         2001             8          21         7.6          22381
## 4         2001            11           6         7.2          35096
## 5         2005             4          19         5.1           2386
## 6         2010             4          20         7.8            333
##    critics_rating critics_score audience_rating audience_score
## 1          Rotten            45         Upright             73
## 2 Certified Fresh            96         Upright             81
## 3 Certified Fresh            91         Upright             91
## 4 Certified Fresh            80         Upright             76
## 5          Rotten            33         Spilled             27
## 6           Fresh            91         Upright             86
##   best_pic_nom best_pic_win best_actor_win best_actress_win best_dir_win
```

```
## 1           no           no             no               no           no
## 2           no           no             no               no           no
## 3           no           no             no               no           no
## 4           no           no            yes               no          yes
## 5           no           no             no               no           no
## 6           no           no             no               no           no
##   top200_box          director            actor1             actor2
## 1         no  Michael D. Olmos    Gina Rodriguez       Jenni Rivera
## 2         no         Rob Sitch         Sam Neill   Kevin Harrington
## 3         no Christopher Guest Christopher Guest   Catherine O'Hara
## 4         no   Martin Scorsese  Daniel Day-Lewis  Michelle Pfeiffer
## 5         no       Stevan Mena     Samantha Dark R. Brandon Johnson
## 6         no   Chung-ryoul Lee     Choi Won-kyun       Lee Sam-soon
##                 actor3           actor4              actor5
## 1 Lou Diamond Phillips    Emilio Rivera Joseph Julian Soria
## 2    Patrick Warburton         Tom Long      Genevieve Mooy
## 3         Parker Posey      Eugene Levy         Bob Balaban
## 4         Winona Ryder Richard E. Grant        Alec McCowen
## 5      Brandon Johnson    Heather Magee      Richard Glover
## 6                  Moo             <NA>               <NA>
##                             imdb_url
## 1 http://www.imdb.com/title/tt1869425/
## 2 http://www.imdb.com/title/tt0205873/
## 3 http://www.imdb.com/title/tt0118111/
## 4 http://www.imdb.com/title/tt0106226/
## 5 http://www.imdb.com/title/tt0388230/
## 6 http://www.imdb.com/title/tt1334549/
##                                       rt_url
## 1     //www.rottentomatoes.com/m/filly_brown_2012/
## 2                 //www.rottentomatoes.com/m/dish/
## 3  //www.rottentomatoes.com/m/waiting_for_guffman/
## 4     //www.rottentomatoes.com/m/age_of_innocence/
## 5 //www.rottentomatoes.com/m/10004684-malevolence/
## 6          //www.rottentomatoes.com/m/old-partner/
```

**Categorical Variables:** title_type, genre, mpaa_rating, critics_rating, audience_rating, best_pic_nom, best_pic_win, best_actor_win, best_actress_win, best_dir_win, top200_box

**Continuous Variables:** runtime, thtr_rel_year, thtr_rel_month, thtr_rel_day, dvd_rel_year, dvd_rel_month, dvd_rel_day, imdb_rating, imdb_num_votes, critics_score, audience_score

**Other Variables:** title, studio, director, actor1, actor2, actor3, actor4, actor5, imdb_url, rt_url

## PART 2: RESEARCH QUESTION

From the data, we can look at possible research questions which are as follows.

- identify the variables influencing imdb_rating

- identify the variables influencing critics_score
- identify the variables influencing audience_score

Both imdb_rating and audience_score for a movie is determined from thousands of votes and could be influenced by biased ratings/scores or even manipulated by bots. Therefore, the critics_score determined from just a few hundred acknowledged critics seems to be the ideal outcome variable for this dataset.

**Identify the predictor variables influencing the outcome variable 'critics_score'.**

## Subsetting Data

The variables 'imdb_url', 'rt_url', 'thtr_rel_year', 'thtr_rel_month', 'thtr_rel_day', 'dvd_rel_year', 'dvd_rel_month', 'dvd_rel_day' are obviously irrelevant in influencing the 'critics_score'.

The variables 'studio', 'director', 'actor1', 'actor2', 'actor3', 'actor4', 'actor5' are just couple of departments in making a movie. Since we don't have any information about other departments, it would be biased to consider these variables in our modeling.

The variables 'best_pic_nom', 'best_pic_win', 'best_actor_win', 'best_actress_win', 'best_dir_win' are acknowledgements announced at the end of an year and cannot influence the 'critics_score' as it is determined almost within the month of a movie release.

The critics are expected to provide their score devoid of any influence from other common movie-goers and market trends. Therefore, the variables such as 'imdb_rating', 'imdb_num_votes', 'audience_score' and 'top200_box' should not be considered.

The variable 'critics_rating' should also not be included in our modeling as it is another outcome variable provided by critics along with the 'critics_score'.

```
sub_movies1 <- subset(movies, select = c(title, critics_score, runtime,
title_type, genre, mpaa_rating))
head(sub_movies1)
```

```
##                   title critics_score runtime   title_type       genre
## 1          Filly Brown            45      80 Feature Film       Drama
## 2             The Dish            96     101 Feature Film       Drama
## 3  Waiting for Guffman            91      84 Feature Film      Comedy
## 4 The Age of Innocence            80     139 Feature Film       Drama
## 5          Malevolence            33      90 Feature Film      Horror
## 6          Old Partner            91      78  Documentary Documentary
##   mpaa_rating
## 1           R
## 2       PG-13
## 3           R
## 4          PG
## 5           R
## 6     Unrated
```

## Missing Values

Find missing values in the data.

```
sub_movies1[!complete.cases(sub_movies1),]
```

```
##                  title critics_score runtime  title_type       genre
## 334 The End of America            80      NA Documentary Documentary
##     mpaa_rating
## 334     Unrated
```

Update the original runtime of the above movie. (www.imdb.com/title/tt1294790/)

```
sub_movies1$runtime[sub_movies1$title=='The End of America'] <- 74
sub_movies1[334,]
```

```
##                  title critics_score runtime  title_type       genre
## 334 The End of America            80      74 Documentary Documentary
##     mpaa_rating
## 334     Unrated
```

# PART 3: EDA

## Summary Statistics

```
sub_movies1 <- subset(sub_movies1, select = -c(title))
sapply(sub_movies1, class)
```

```
## critics_score       runtime    title_type         genre   mpaa_rating
##     "numeric"     "numeric"      "factor"      "factor"      "factor"
```

```
summary(sub_movies1)
```

```
##  critics_score       runtime             title_type
##  Min.   :  1.00   Min.   : 39.0   Documentary : 55
##  1st Qu.: 33.00   1st Qu.: 92.0   Feature Film:591
##  Median : 61.00   Median :103.0   TV Movie    :  5
##  Mean   : 57.69   Mean   :105.8
##  3rd Qu.: 83.00   3rd Qu.:115.5
##  Max.   :100.00   Max.   :267.0
##
##                 genre      mpaa_rating
##  Drama             :305   G      : 19
##  Comedy            : 87   NC-17  :  2
##  Action & Adventure: 65   PG     :118
##  Mystery & Suspense: 59   PG-13  :133
##  Documentary       : 52   R      :329
##  Horror            : 23   Unrated: 50
##  (Other)           : 60
```

## Categorical Predictor Variables

All of the categorical predictor variables are unbalanced but more importantly, the variables 'title_type' and 'mpaa_rating' require change of reference level.

```
sub_movies1$title_type <- relevel(sub_movies1$title_type, "Feature Film")
table(sub_movies1$title_type)
```

```
##
## Feature Film  Documentary     TV Movie
##          591           55            5
```

```
sub_movies1$mpaa_rating <- relevel(sub_movies1$mpaa_rating, "R")
table(sub_movies1$mpaa_rating)
```

```
##
##       R       G   NC-17      PG   PG-13 Unrated
##     329      19       2     118     133      50
```

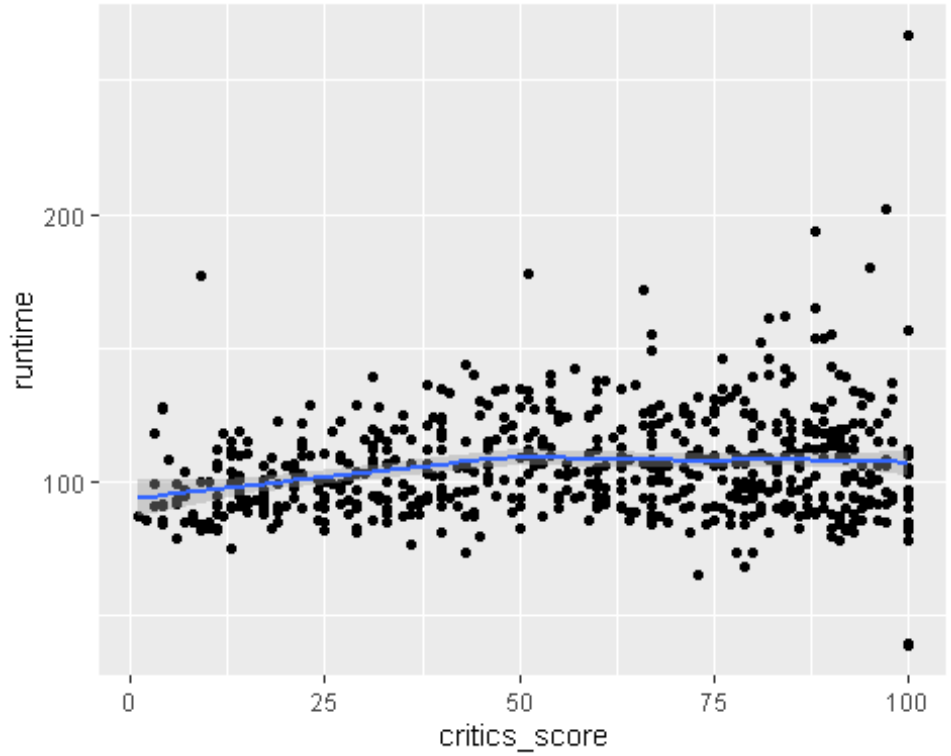
## Continuous Predictor Variables

There is only one continuous predictor variable 'runtime' which is plotted against the outcome variable to check for non-linear relationship.

```
library(ggplot2)

ggplot(data = sub_movies1, aes(critics_score, runtime)) + geom_point() +
geom_smooth()
```

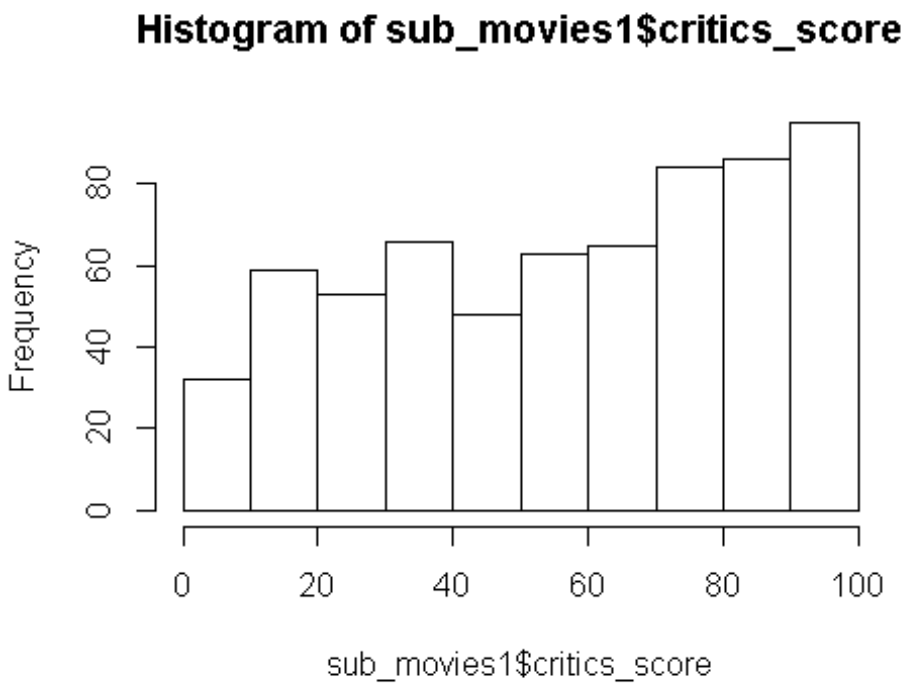
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




The relationship does not appear to be non-linear and/or heteroscedastic and hence a transformation is not required.

## Outcome Variable

```
hist(sub_movies1$critics_score)
```




The histogram of outcome variable does not exhibit a perfect normal distribution but it is also NOT non-normal or non-symmetric or skewed and hence, it does not require any transformation.

# PART 4: MODELING

## Univariate Regression Models

Single regression models are fitted between the outcome variable 'critics_score' and each predictor variable. The resulting p values are checked whether they are less than 0.2 and are included in multivariate regression.

```
library(car)

## Loading required package: carData

model_runtime <- glm(critics_score~runtime, family=gaussian,
data=sub_movies1)
model_title_type <- glm(critics_score~title_type, family=gaussian,
data=sub_movies1)
model_genre <- glm(critics_score~genre, family=gaussian, data=sub_movies1)
model_mpaa_rating <- glm(critics_score~mpaa_rating, family=gaussian,
data=sub_movies1)

Anova(model_runtime)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: critics_score
##         LR Chisq Df Pr(>Chisq)
## runtime   19.335  1  1.097e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(model_title_type)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: critics_score
##            LR Chisq Df Pr(>Chisq)
## title_type   80.413  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(model_genre)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: critics_score
##       LR Chisq Df Pr(>Chisq)
## genre   154.73 10  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(model_mpaa_rating)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: critics_score
##             LR Chisq Df Pr(>Chisq)
## mpaa_rating   70.679  5  7.403e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value of all the predictor variables are less than 0.2 and can be considered for inclusion in multivariate regression.

## Multivariate Regression Models

A multivariate regression model is fitted between the outcome variable and all the predictor variables together and a backward elimination approach is implemented using the 'step' function. The 'step' function fits multiple models between the variables and outputs Akaike Information Criterion (AIC) score of all the models - lesser the AIC score, better the model.

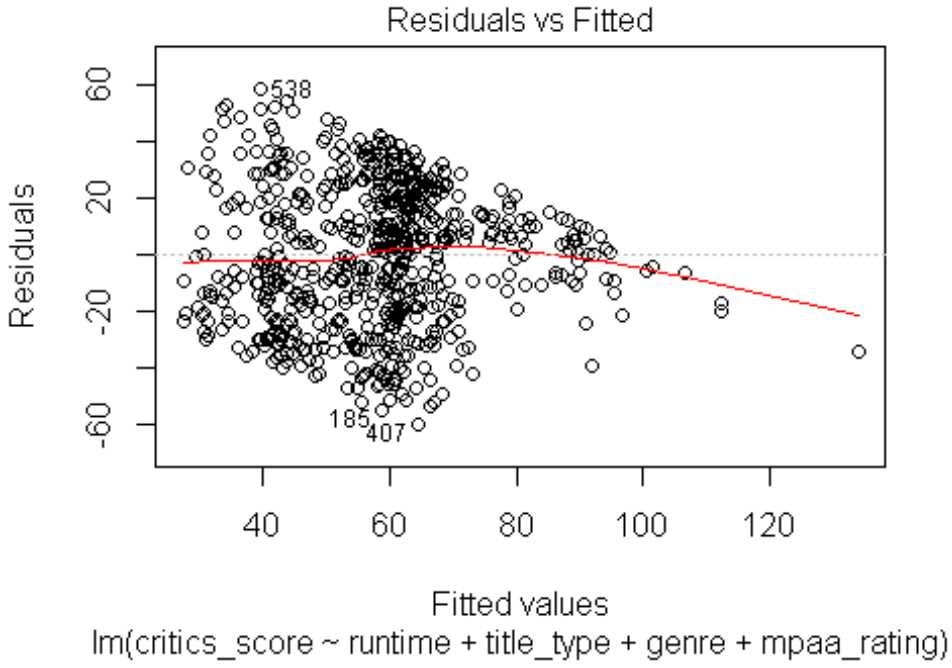
```
full_model <- step(glm(critics_score~runtime+title_type+genre+mpaa_rating,
family=gaussian, data=sub_movies1), direction = "backward")
```

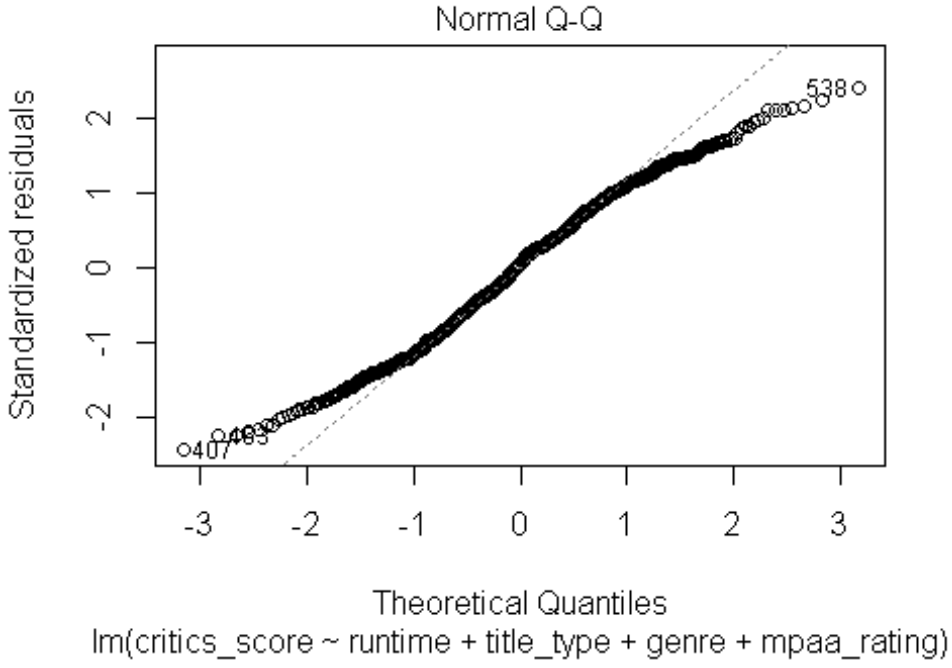
```
## Start:  AIC=6041.88
## critics_score ~ runtime + title_type + genre + mpaa_rating
##
##               Df Deviance    AIC
## <none>             384655 6041.9
## - title_type   2   391587 6049.5
## - runtime      1   397573 6061.4
## - mpaa_rating  5   403575 6063.1
## - genre       10   420119 6079.3
```

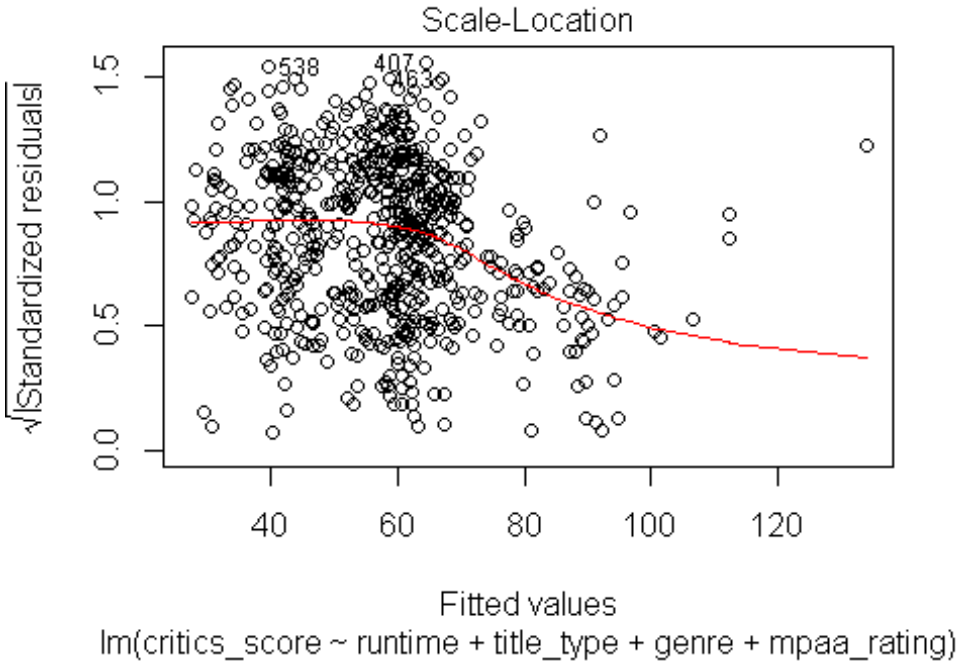
In this case, the 'step' function has output only one model and deems it to be the best model, suggesting not to consider other models with different combinations of predictor variables.

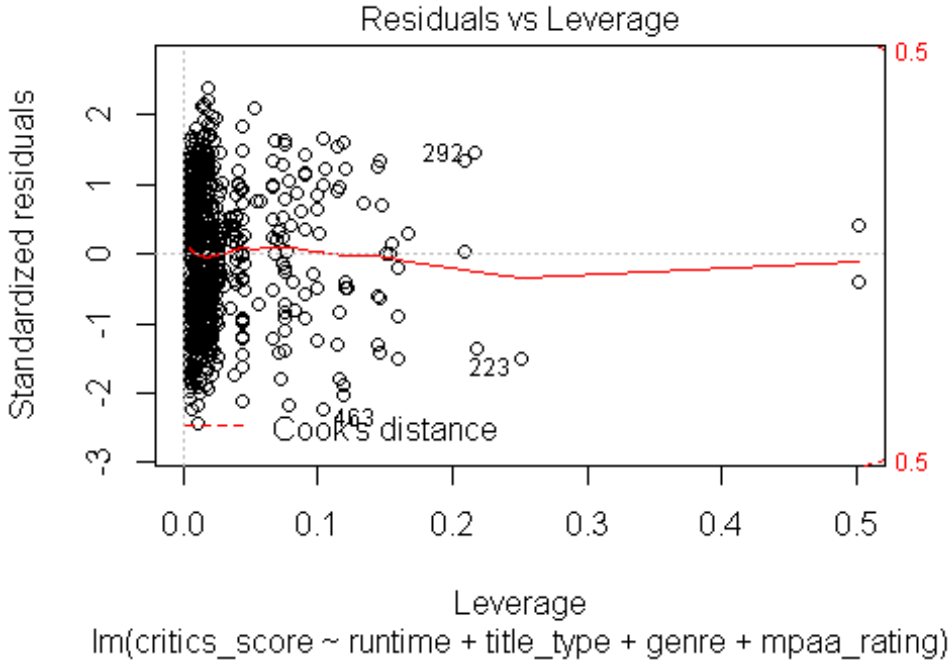
## Model Diagnostics

```
best_model <- lm(critics_score~runtime+title_type+genre+mpaa_rating,
data=sub_movies1)
plot(best_model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(critics_score ~ runtime + title_type + genre + mpaa_rating)
538
185
407
-60
-20
20
60
40
60
80
100
120


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(critics_score ~ runtime + title_type + genre + mpaa_rating)
538
407
-2
-1
0
1
2
-3
-2
-1
0
1
2
3




**Residuals vs Fitted plot:** Barring a few obvious outliers, residuals are in a random scatter largely around the zero line

**Normal Q-Q plot:** Residuals are not exactly lined up on the dashed line but are largely distributed around it, indicating normality

**Scale-Location plot:** Residuals are not spreading wider and wider along the x axis on a steep angle which is ideal

**Residuals vs Leverage plot:** No residuals influencing the regression outside the Cook's distance line which is at the edges

## Model Interpretation

```
summary(best_model)
```

```
##
## Call:
## lm(formula = critics_score ~ runtime + title_type + genre + mpaa_rating,
##     data = sub_movies1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.557 -19.400   1.652  19.448  58.234
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     16.2182     6.5302   2.484 0.013266 *
## runtime                          0.2479     0.0538   4.607 4.94e-06 ***
## title_typeDocumentary           30.4219     9.1428   3.327 0.000927 ***
## title_typeTV Movie              -5.5169    11.3229  -0.487 0.626259
## genreAnimation                  -0.1853     9.6219  -0.019 0.984638
## genreArt House & International   7.6743     7.4443   1.031 0.302985
## genreComedy                      2.6820     4.1070   0.653 0.513975
## genreDocumentary                10.7032     9.8141   1.091 0.275869
## genreDrama                      19.8160     3.4726   5.706 1.78e-08 ***
## genreHorror                      3.8426     6.1491   0.625 0.532255
## genreMusical & Performing Arts  20.7216     8.3694   2.476 0.013552 *
## genreMystery & Suspense         12.5807     4.5522   2.764 0.005882 **
## genreOther                      21.2013     6.9627   3.045 0.002424 **
## genreScience Fiction & Fantasy   7.2127     8.7850   0.821 0.411945
## mpaa_ratingG                    17.9783     6.7109   2.679 0.007577 **
## mpaa_ratingNC-17                22.1791    17.5320   1.265 0.206315
## mpaa_ratingPG                    1.7528     2.7619   0.635 0.525891
## mpaa_ratingPG-13                -9.4262     2.6170  -3.602 0.000341 ***
## mpaa_ratingUnrated              10.4320     4.7632   2.190 0.028879 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.67 on 632 degrees of freedom
## Multiple R-squared:  0.2664, Adjusted R-squared:  0.2456
## F-statistic: 12.75 on 18 and 632 DF,  p-value: < 2.2e-16
```

A movie has a high probability of lower 'critics_score' when the movie is a TV movie or Animation movie or PG-13 movie.

A movie has a high probability of higher 'critics_score' when the movie falls in the below categories.

- genre: Art House & International, Comedy, Documentary, Drama, Horror, Musical & Performing Arts, Mystery & Suspense, Science Fiction & Fanatsy, Other
- mpaa_rating: G, NC-17, PG, Unrated

## PART 5: PREDICTION

The movie 'Parasite' (www.rottentomatoes.com/m/parasite_2019) recently nominated for Oscars 2020 is considered for analyzing the predictive ability of the model.

```
new_movie_parasite <- data.frame(runtime = 132, title_type = "Feature Film" ,
genre = "Drama" , mpaa_rating = "R")

predict(best_model, newdata = new_movie_parasite)

##        1
## 68.75271
```

The original 'critics_score' of the movie is 99 and the predicted value falls well short of it.

```
predict(best_model, newdata = new_movie_parasite, interval = "confidence")

##        fit      lwr      upr
## 1 68.75271 64.74329 72.76214

predict(best_model, newdata = new_movie_parasite, interval = "prediction")

##        fit      lwr      upr
## 1 68.75271 20.14108 117.3643
```

As we are predicting for a specific individual movie, the prediction interval is more appropriate to use, despite the higher interval of 117 is not possible with the maximum 'critics_score' being 100.

## PART 6: CONCLUSION

The variables 'runtime', 'title_type', 'genre' and 'mpaa_rating' are identified to be influencing the outcome variable 'critics_score'.

In hindsight, it would be interesting to see the results when more variables are considered for modelling at the start. It could have probably produced a much better prediction.