Modeling and prediction for movies

Linear Regression and Modeling - Statistics with R - Duke University

Pedro M. Gallardo

April '17

Setup

Load packages

```
library(ggplot2)
library(dplyr)
library(statsr)
library(gridExtra)
library(knitr)
```

Load data

```
load("movies.Rdata")
```

Part 1: Data

The data set is comprised of 651 randomly sampled movies produced and released before 2016.

One of the main problems that might have the linear model is collinearity.

Part 2: Research question

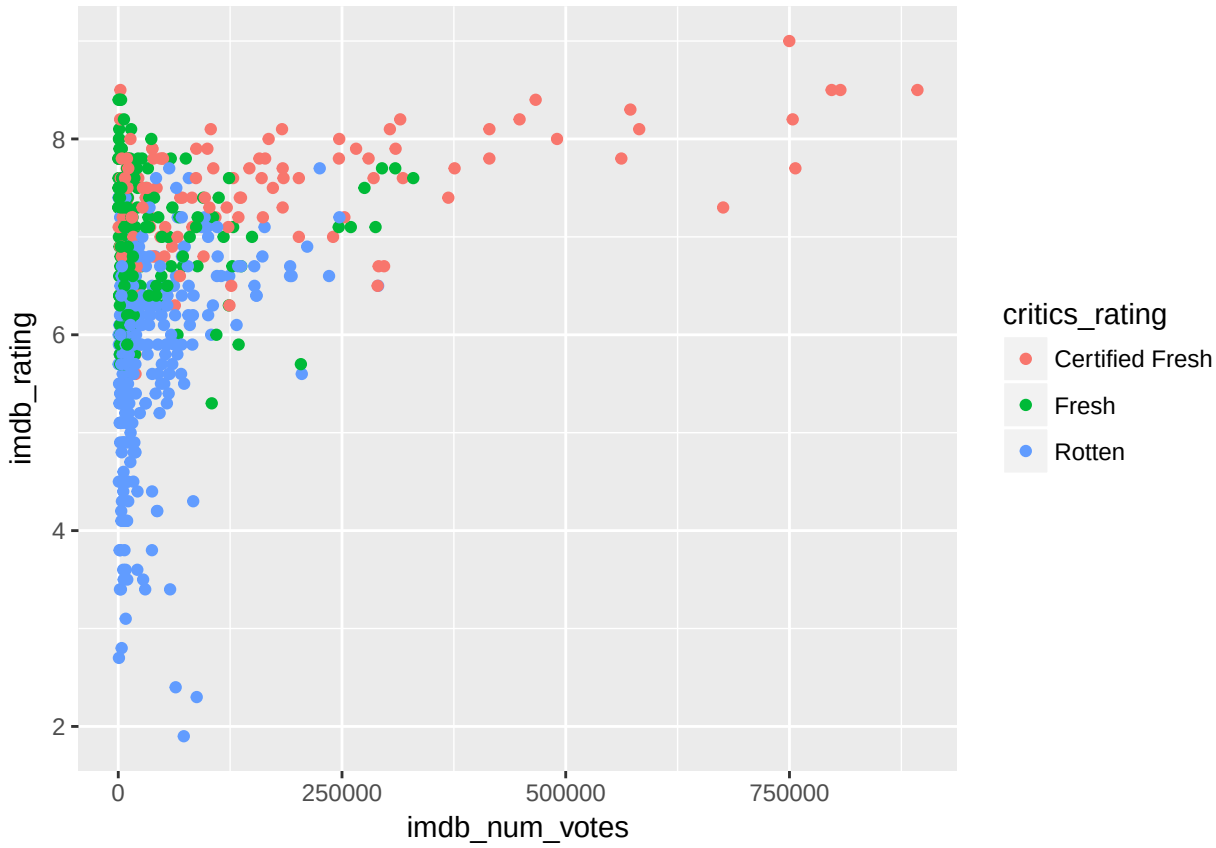
The research question that is going to be addressed is: **What attributes make a movie popular?**.

Part 3: Exploratory data analysis

First of all, this analysis is going to define what is understood as "popular".

It is of the researcher understanding that the rating of a movie along with the number of people who has watched it makes it "popular". Therefore, a new variable that combines these two variables from IMDB, **imbd_rating** and **imbd_num_votes** is going to be created so as to be the response variable of this study.

```
ggplot(data=movies, aes(x=imdb_num_votes ,y=imdb_rating, colour=critics_rating ))+geom_point()
```




```
movies$success<-movies$imdb_rating*log(movies$imdb_num_votes)
```

The new variable is called **success**, it is a numerical variable and it is computed as: **success** = **imbd_rating** X log(**imbd_num_votes**)

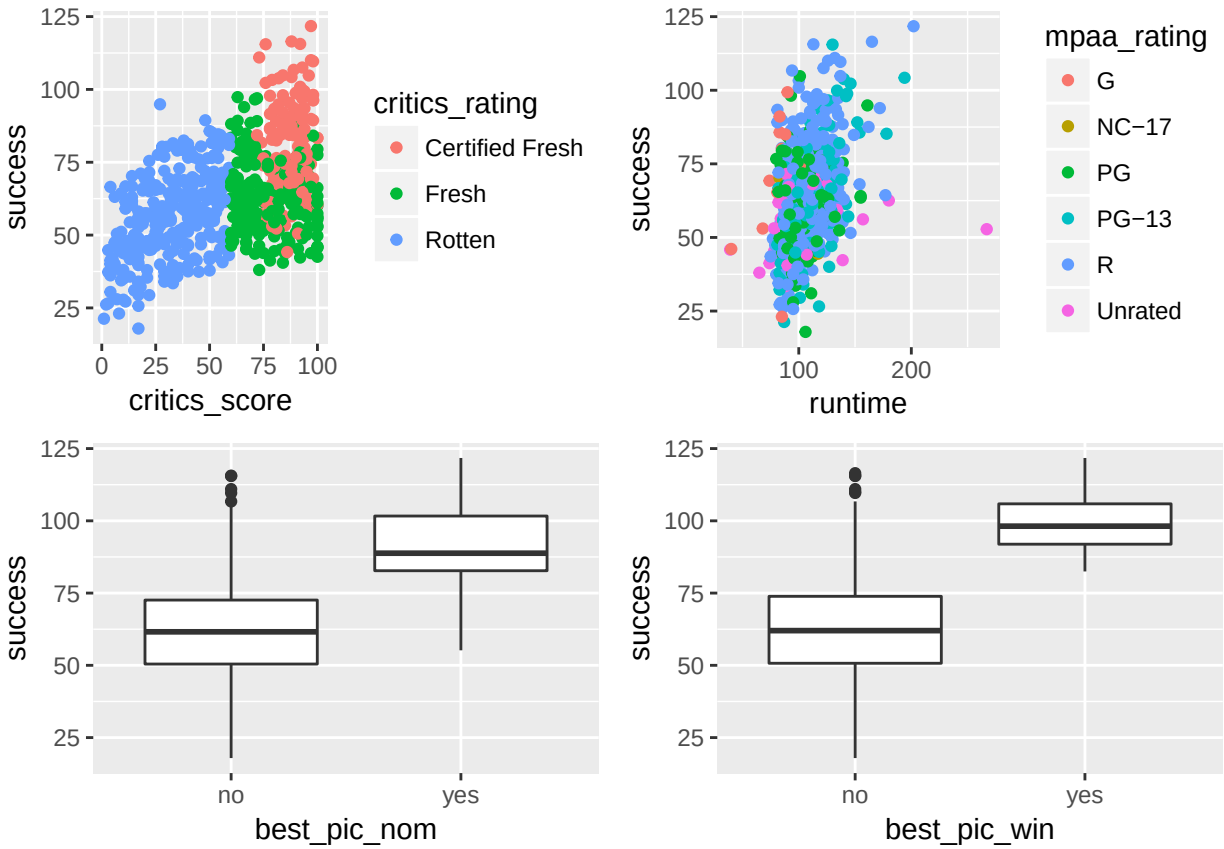
It has been used the logarithm of the **imbd_num_votes** in order to account for the order of the number of voters rather than the number as itself so as to have a response variable within a smaller range.

Next, the response varible is going to be plotted against some of the variables that are going to be used as explanatory variables.

```
par(mfrow=c(2,2))
a<-ggplot(data=movies, aes(x=critics_score ,y=success, colour=critics_rating))+geom_point()
b<- ggplot(data=movies, aes(x=runtime ,y=success, colour=mpaa_rating ))+geom_point()
c<- ggplot(data=movies, aes(x=best_pic_nom ,y=success))+geom_boxplot()
d<- ggplot(data=movies, aes(x=best_pic_win ,y=success))+geom_boxplot()

 grid.arrange(a,b,c,d)
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



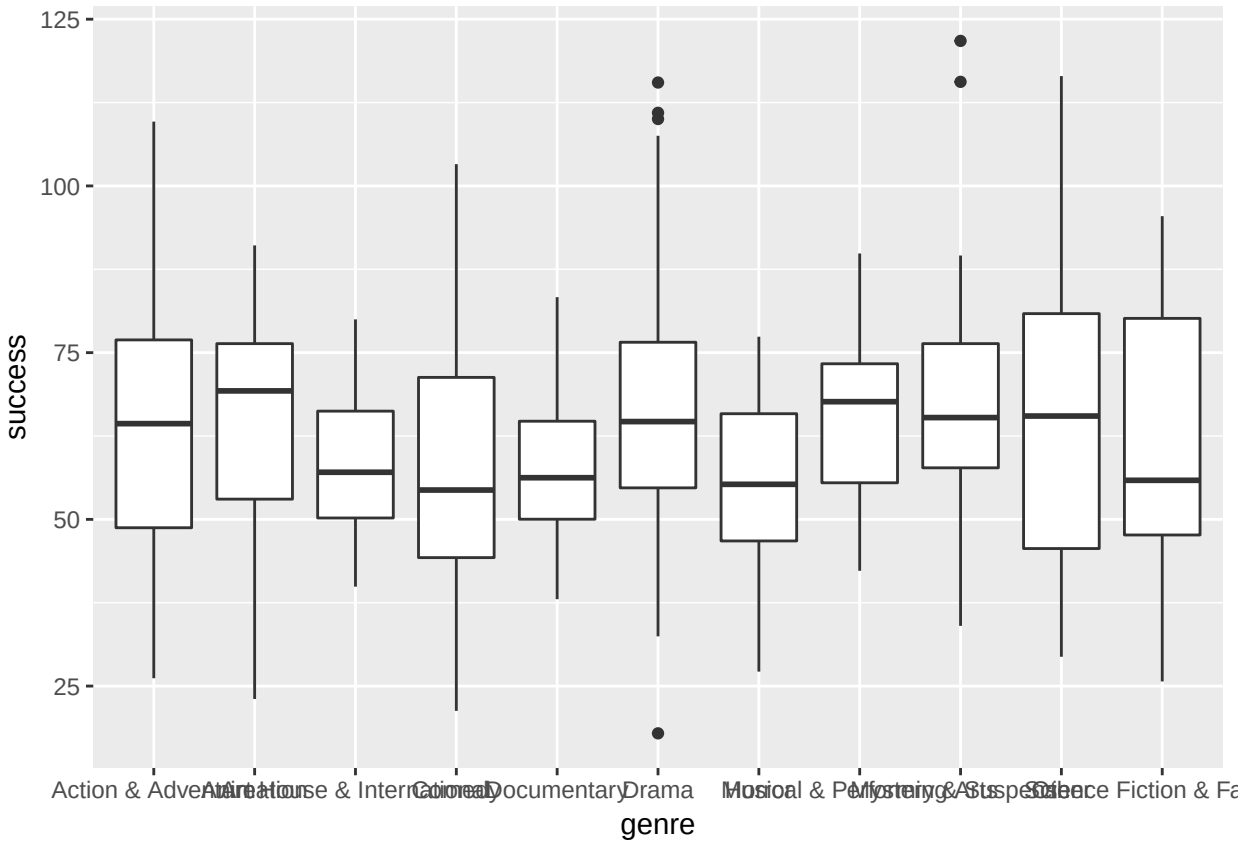

```
 a<-movies%>%
  group_by(genre) %>%
  summarise(medianSuccess=median(success), medianNumber=median(imdb_num_votes), medianRating=median(imdb
arrange(desc(medianSuccess))

 a
```

```
## # A tibble: 11 × 6
##                        genre medianSuccess medianNumber medianRating
##                       <fctr>         <dbl>        <dbl>        <dbl>
## 1                  Animation      69.26499      54363.0         6.40
## 2  Musical & Performing Arts      67.64315      10550.5         7.55
## 3                      Other      65.49524      16371.0         6.80
## 4          Mystery & Suspense      65.26285      25264.0         6.50
## 5                      Drama      64.66209      15025.0         6.80
## 6          Action & Adventure      64.35563      48718.0         6.00
## 7  Art House & International      57.06576       5812.5         6.50
## 8                Documentary      56.24560       1791.5         7.60
## 9  Science Fiction & Fantasy      55.85554      13790.0         5.90
## 10                    Horror      55.25566      16824.0         5.90
## 11                    Comedy      54.39861      14949.0         5.70
## # ... with 2 more variables: IQR <dbl>, count <int>
```

```
b<-movies%>%
group_by(genre)

ggplot(b, aes(y=success, x=genre))+geom_boxplot()
```



From previous graphs, it may be observed that:

- the values for **success** variable ranges among 0 - 125, it does not have a unit, it is just and index. The higher it is, the most popular the film is.
- **critics_score**, **best_pic_nom**, **best_pic_win** seem to have a clear correlation with the "popularity" of the film.
- From the film **genre** exploration, it is observed how works the **success** variable, despides the documentaries rates are high, they are barely seen by the people, on the contrary, the animation films, are seen by a higher number of people but do not score that high. The balance between film score and amount of viewers leans the balance towards viewers. So in this exploratory analysis is concluded that the animation films are the most popular films.

Part 4: Modeling

The method to build the model is going to be the backwards elimination and the selection criterion is going to be the adjusted R^2.

In the model, the response variable is the **success** variable and the explanatory variables are all the following variables:

1. critics_score
2. best_pic_nom
3. best_pic_win
4. best_actor_win
5. best_actress_win
6. best_dir_win

7. genre
8. mpaa_rating
9. runtime
10. top200_box

Variables such as **critics_rating**, **audicence_rating**, **audience_score** have been eliminated from the model to avoid collinearity and **director** and **actor1** have been eliminated as when adding them the conditions for the MLR model are not met.

Step 1:

```
model0<- lm(success ~ critics_score + best_pic_nom + best_pic_win + best_actor_win + best_actress_win +

model1<-lm(success ~ best_pic_nom + best_pic_win + best_actor_win + best_actress_win +  best_dir_win + g

model2<- lm(success ~ critics_score + best_pic_win + best_actor_win + best_actress_win +  best_dir_win

model3<- lm(success ~ critics_score + best_pic_nom + best_actor_win + best_actress_win +  best_dir_win

model4<- lm(success ~ critics_score + best_pic_nom + best_pic_win + best_actress_win +  best_dir_win + g

model5<- lm(success ~ critics_score + best_pic_nom + best_pic_win + best_actor_win + best_dir_win + geni

model6<- lm(success ~ critics_score + best_pic_nom + best_pic_win + best_actor_win + best_actress_win +

model7<- lm(success ~ critics_score + best_pic_nom + best_pic_win + best_actor_win + best_actress_win +

model8<- lm(success ~ critics_score + best_pic_nom + best_pic_win + best_actor_win + best_actress_win +

model9<- lm(success ~ critics_score + best_pic_nom + best_pic_win + best_actor_win + best_actress_win +

model10<- lm(success ~ critics_score + best_pic_nom + best_pic_win + best_actor_win + best_actress_win

kable(data.frame(variable_eliminated=c("None", "critics_score"," best_pic_nom", "best_pic_win", "best_a
```

variable_eliminated	Rs
None	0.4373552
critics_score	0.2203816
best_pic_nom	0.4301016
best_pic_win	0.4370081
best_actor_win	0.4381141
best_actress_win	0.4382361
best_dir_win	0.4366803
genre	0.4262097
mpaa_rating	0.4122695
runtime	0.3856867
top200_box	0.4037658

From this first step, the variable **best_actress_win** and **best_actor_win** are eliminated.

Step 2

```
model0<- lm(success ~ critics_score + best_pic_nom + best_pic_win + best_dir_win + genre + mpaa_rating +

model1<-lm(success ~ best_pic_nom + best_pic_win + best_dir_win + genre + mpaa_rating + runtime + top20

model2<- lm(success ~ critics_score + best_pic_win + best_dir_win + genre + mpaa_rating + runtime + top

model3<- lm(success ~ critics_score + best_pic_nom +  best_dir_win + genre + mpaa_rating + runtime + to

model4<- lm(success ~ critics_score + best_pic_nom + best_pic_win +  genre + mpaa_rating + runtime + to

model5<- lm(success ~ critics_score + best_pic_nom + best_pic_win + best_dir_win + mpaa_rating + runtim

model6<- lm(success ~ critics_score + best_pic_nom + best_pic_win + best_dir_win + genre + runtime + to

model7<- lm(success ~ critics_score + best_pic_nom + best_pic_win + best_dir_win + genre + top200_box, d

model8<- lm(success ~ critics_score + best_pic_nom + best_pic_win + best_dir_win + genre + runtime, data

kable(data.frame(variable_eliminated=c("None", "critics_score"," best_pic_nom", "best_pic_win", "best_di
```

variable_eliminated	Rs
None	0.4389973
critics_score	0.2228262
best_pic_nom	0.4313897
best_pic_win	0.4387157
best_dir_win	0.4382748
genre	0.4277027
mpaa_rating	0.4140870
runtime	0.3859214
top200_box	0.4055765
“‘	

This step is the last one. The parsimony model has been reached. When any other variable is removed from the model the adjusted-R^2 diminishes.

So, the final model is:

```
summary(model0)
```

```
##
## Call:
## lm(formula = success ~ critics_score + best_pic_nom + best_pic_win +
##     best_dir_win + genre + mpaa_rating + runtime + top200_box,
##     data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.487  -8.832  -0.723   8.458  40.666
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
```

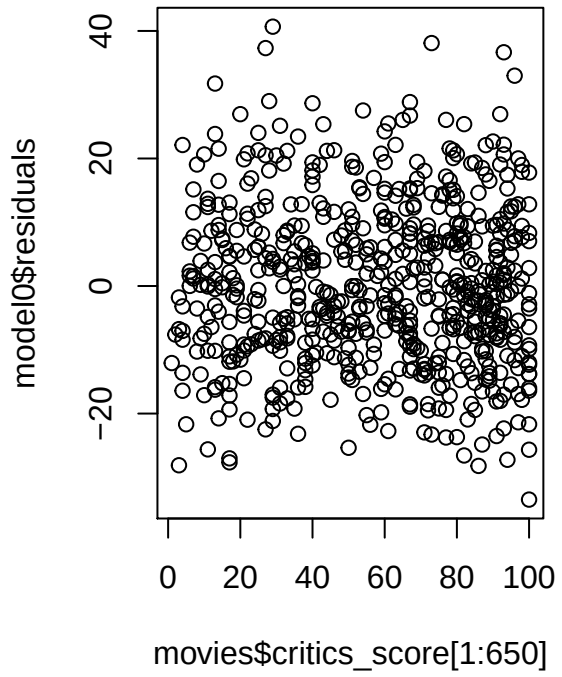
```
## (Intercept)                     30.43794    4.56344   6.670 5.62e-11 ***
## critics_score                    0.32890    0.02108  15.600  < 2e-16 ***
## best_pic_nomyes                 10.17392    3.29572   3.087 0.002111 **
## best_pic_winyes                  6.69220    5.83417   1.147 0.251789
## best_dir_winyes                  2.98130    2.21590   1.345 0.178977
## genreAnimation                   1.85300    5.02493   0.369 0.712429
## genreArt House & International  -5.73817    3.87870  -1.479 0.139533
## genreComedy                     -5.23264    2.14373  -2.441 0.014926 *
## genreDocumentary               -12.17776    2.91926  -4.172 3.45e-05 ***
## genreDrama                      -5.78502    1.86450  -3.103 0.002004 **
## genreHorror                     -6.12966    3.20291  -1.914 0.056103 .
## genreMusical & Performing Arts  -9.07945    4.13779  -2.194 0.028581 *
## genreMystery & Suspense         -2.87619    2.38982  -1.204 0.229230
## genreOther                      -4.72844    3.64560  -1.297 0.195099
## genreScience Fiction & Fantasy  -6.21885    4.57290  -1.360 0.174338
## mpaa_ratingNC-17               -10.81451    9.72544  -1.112 0.266572
## mpaa_ratingPG                   -1.84540    3.55123  -0.520 0.603489
## mpaa_ratingPG-13                 4.13857    3.65975   1.131 0.258556
## mpaa_ratingR                     2.83672    3.53549   0.802 0.422651
## mpaa_ratingUnrated              -8.08472    4.03159  -2.005 0.045355 *
## runtime                          0.16192    0.02992   5.412 8.88e-08 ***
## top200_boxyes                   11.77872    3.49185   3.373 0.000789 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.83 on 628 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4571, Adjusted R-squared:  0.439
## F-statistic: 25.18 on 21 and 628 DF,  p-value: < 2.2e-16
```

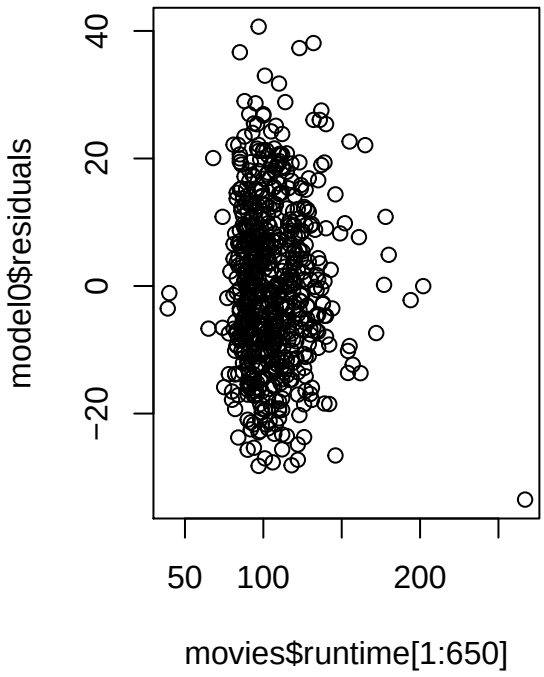
A quick check for the multiple linear regression condition is going to be conducted to ensure the validity of the model:

MLR conditions:

Linear relationships between x and y

```
par(mfrow=c(1,2))
plot(model0$residuals ~ movies$critics_score[1:650])
plot(model0$residuals ~ movies$runtime[1:650])
```

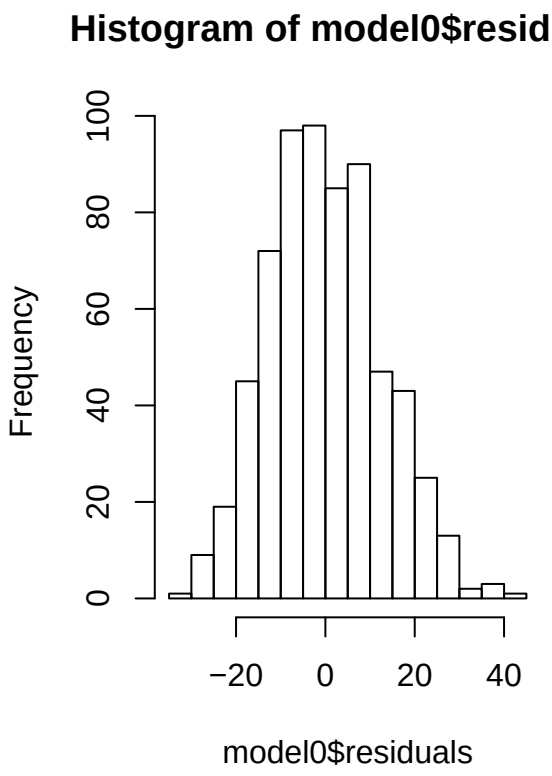


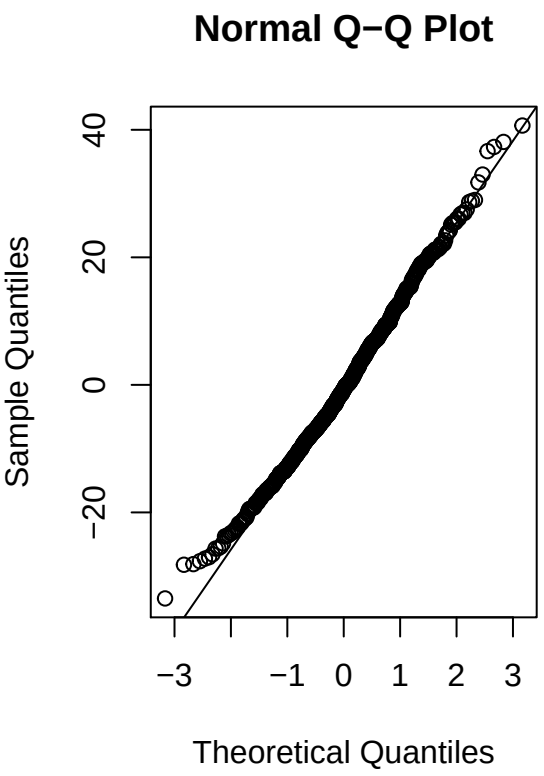


It is only check for the numerical variables. The condition is met. As the variables show a linearity with the response variable.

Nearly normal residuals

```
par(mfrow=c(1,2))
hist(model0$residuals)
qqnorm(model0$residuals)
qqline(model0$residuals)
```

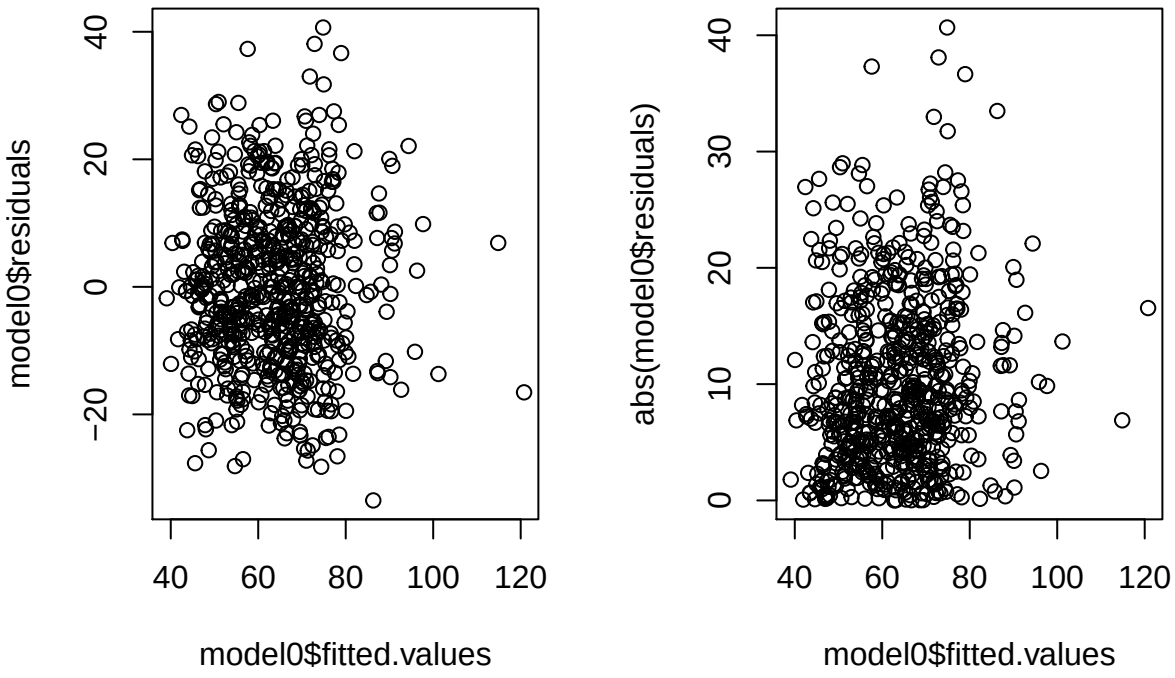




The condition is also met.

Constant variability of residuals

```
par(mfrow=c(1,2))
plot(model0$residuals ~ model0$fitted.values)
plot(abs(model0$residuals) ~ model0$fitted.values)
```



Not fan shape is observed hence the condition is met.

Independece of residuals

The data have been randomly sampled as it was stated at the beginning of the report. Therefore the condition is also met.

Part 5: Prediction

The film chosen to be predicted is Rogue one by Gareth Edwards, prologue of the famous stars wars saga. These are variables introduced in the model, all of them have been taken from imdb website:

Film: Rogue one: A Star Wars Story. Critics_score: 65 (from imdb) best_pic_nom= no, best_pic_win= no, best_dir_win= no, genre="Action & Adventure", mpaa_rating="PG-13", runtime= 133 minutes, top200_box= no

```
new=data.frame(critics_score=65, best_pic_nom="no", best_pic_win="no",best_dir_win="no", genre="Action &

predict(model0, newdata=new, interval="prediction")
```

```
##        fit      lwr      upr
## 1 77.49106 51.95043 103.0317
```

```
#
real_value<-log(282751)*8
real_value
```

```
## [1] 100.4186
```

Conclusion on the prediction

The real value is within the predicted range, although it is close to the upper limit. This might be due to the popularity of the saga (Star Wars) that is raising the popularity of the film and it is an aspect that is not covered by the model herein built.

Part 6: Conclusion

The key to the success of a film lays in the difussion it may get. The oscar nomination is a signigicant booster of its popularity.

According to the Multiple Linear Regression model if a movie is nominated for the oscars or/and wins it, the popularity of the movie is going to raise significantly (10,17 and 6,69 respectively).

The popularity variable herein study is not driven by the fame of the actors/actresses that perfom in the film and the genre of the film only is significantly relevant (i.e: p-value <0.05) for the genres of documentary, comedy, drama and musical. It is interesting to note that despide the documentaries usually score high they are not very popular, most probably because they are films for niches.