Modeling and prediction for movies

Raya Bhattacharya

23 Dec 2017

Load packages

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(statsr)
```

```
## Warning: package 'statsr' was built under R version 3.4.4
```

```
## Loading required package: BayesFactor
```

```
## Warning: package 'BayesFactor' was built under R version 3.4.4
```

```
## Loading required package: coda
```

```
## Warning: package 'coda' was built under R version 3.4.4
```

```
## Loading required package: Matrix
```

```
## ************
## Welcome to BayesFactor 0.9.12-4.2. If you have questions, please contact
Richard Morey (richarddmorey@gmail.com).
##
## Type BFManual() to open the manual.
## ************
```

Load data

```
load("C:/Users/raya.bhattacharya/Downloads/R Data/movies.rdata")
```

Part 1: Data

The data used here is about how much audiences and critics like movies as well as numerous other variables about the movies. Random Sampling has been used to collect the data from random poeple about their observation.

Random sample of movies were taken, so the movies can be said to be independent and the results are generalizable for the whole population.

But the situation does not consist of an experiment being rather can be data is observational(from Rotten Tomatoes & IMDB). So the results which we are going to obtain from our analysis below can't be used to establish a causal relationship.

Part 2: Research question

To find out what attributes make a movie popular I would like to find the associativity between movie attributes and its reception score in general.I will be exploring the association between reception score from multiple sources(IMDB,Rotten Tomatoes,critics reviews) considering the attributes such as movie's cast,director,production studio,runtime & genre.

The reason why I chose the above factors because I feel the analysis based on the above factors woul prove to be of interest for movies industry as the answer to the question will potentially lead to the understanding of factors which makes a movie popular & successful.

Part 3: Exploratory data analysis

To provide a better understanding I would perform my research only on "Non-Documentary" & "Non-Tv movies".

```
#Creating a new data set required_movies consist of filtered movies based on
title_type & genre

required_movies<- movies %>% filter(movies$title_type == "Feature Film",
genre!="Documentary")
```

Now we want to calculate the average score based on the "imdb rating", "critics score","audience score".

But we have a problem regarding the scale of rating. Imdb rates on a scale of 10 and the other two rates on a scale of 100.

So we do the below conversion and find the average score:

```
required_movies$avg_score <- (required_movies$imdb_rating*10 +
required_movies$critics_score + required_movies$audience_score)/3
```

Now we look into the variable studio in the dataset required_movies to explore the average movie score across different studios and look which studios have the highest score in the sample:

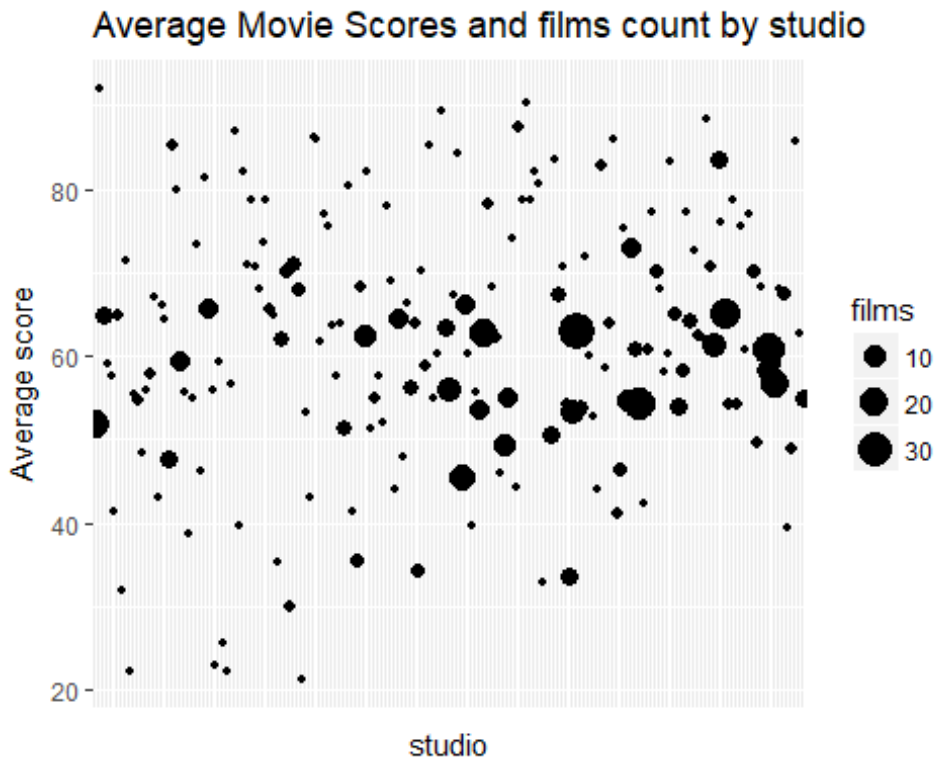
```
stud <- required_movies %>% group_by(studio) %>% summarise(score =
mean(avg_score), films = n()) %>% arrange(-score)
stud
```

```
## # A tibble: 183 x 3
##                            studio    score films
##                            <fctr>    <dbl> <int>
##  1 20th Century Fox Film Corporat 92.00000     1
##  2                Newmarket Films 90.33333     1
##  3       Magnet/Magnolia Pictures 89.33333     1
##  4                        TriStar 88.33333     1
##  5               New Yorker Films 87.33333     2
##  6                         Disney 87.00000     1
##  7                   Good Machine 86.33333     1
##  8              Gramercy Pictures 86.00000     1
##  9         Samuel Goldwyn Company 86.00000     1
## 10                        Winstar 85.66667     1
## # ... with 173 more rows
```

It can be seen that the 20th century Fox Film Corporation has the highest average score of 92.

Below is the graphical representation of the average scores of the studio and the number of films produced by the corresponding studio.

```
ggplot(data = stud, aes(x=studio, y=score)) + geom_point(aes(size=films)) +
theme(axis.text.x=element_blank(),axis.ticks.x= element_blank()) +
ggtitle("Average Movie Scores and films count by studio") + xlab("studio") +
ylab("Average score")
```

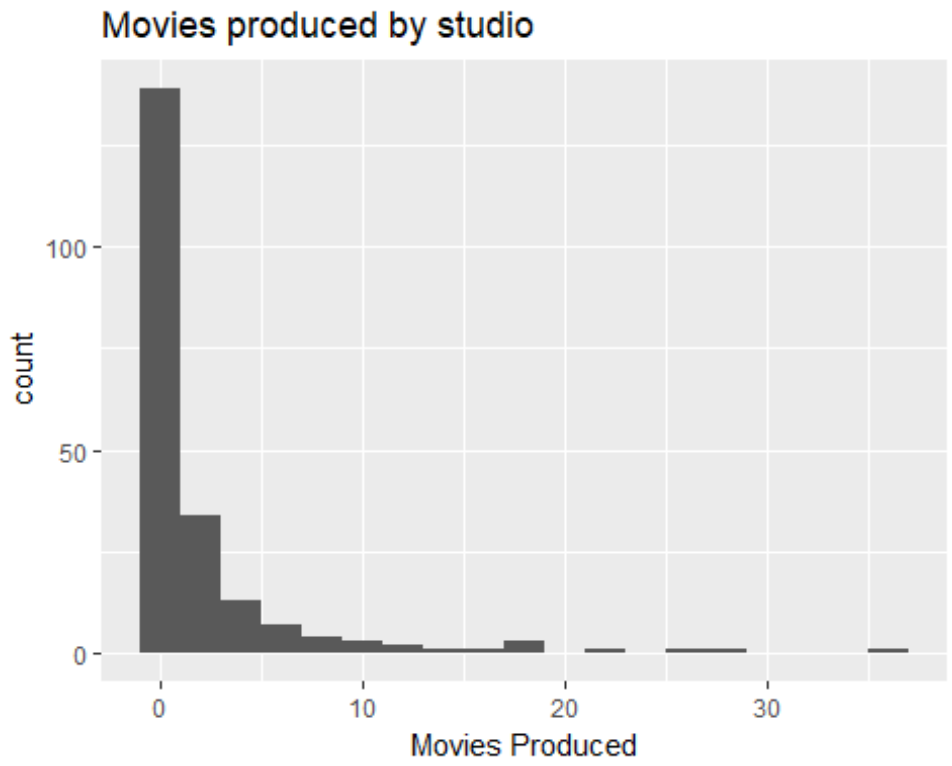


What we see in the above graphical representation?

The Big studios which have produced many films does not have high average scores.But this seems to be just due to the averaging scores from many films, not an actual pattern in the data. I am still going to include data about studios in my further research but with a transformed version of "studio" into "big_studio" which represents whether a film was produced by a big studio.

To determine a studio as a big studio we need to check the films produced by each studio which is done is the below code:

```
studio_df <- data.frame(table(required_movies$studio))
ggplot(data = studio_df,aes(x=Freq)) + geom_histogram(binwidth = 2) +
ggtitle("Movies produced by studio") + xlab("Movies Produced")
```



Now we add the new variable big_studio for the studios which has produced 5 or more movies

```
big_studio<- as.character(studio_df[studio_df$Freq>=5,]$Var1)
required_movies$big_studio <- ifelse(required_movies$studio %in% big_studio,
TRUE, FALSE)
```

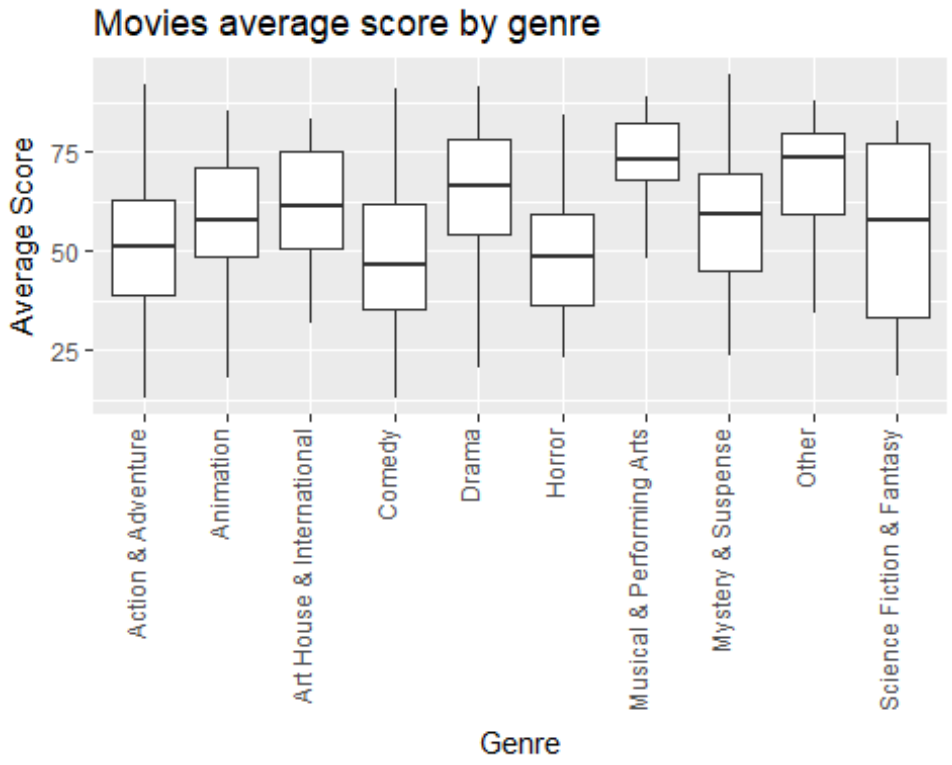
Finally we will be exploring “genre” and its association with “avg_score”.

```
required_movies %>% group_by(genre) %>% summarise(score = mean(avg_score),
films = n()) %>% arrange(-score)
```

```
## # A tibble: 10 x 3
##                         genre    score films
##                        <fctr>    <dbl> <int>
##  1 Musical & Performing Arts 72.83333     8
##  2                     Other 68.55556    15
##  3                     Drama 64.73533   301
##  4 Art House & International 60.59524    14
##  5        Mystery & Suspense 58.55932    59
##  6                 Animation 57.22222     9
##  7 Science Fiction & Fantasy 52.81481     9
##  8        Action & Adventure 51.64103    65
##  9                    Comedy 49.74510    85
## 10                    Horror 49.13043    23
```

Now we see that the genres 'Drama' and 'Musical & Performing Arts' have highest average scores. The boxplot below displays range of scores by genre.

```
ggplot(data=required_movies,aes(x=genre,y=avg_score)) + geom_boxplot() +
theme(axis.text.x= element_text(angle = 90,hjust = 1,vjust = 0)) +
ggtitle("Movies average score by genre") + xlab("Genre") + ylab("Average
Score")
```



Part 4: Modeling

Here we will develop a linear regression model to predict the average movie score (avg_score variable created in previous section).

From initial set of variables presented in the dataset only the following will be left as predictors: "genre", "big studio" , "runtime", "mpaa rating", "best actor win", "best actress win", "best dir win".

The other variables were initially excluded from the modelling, because intuitively they are not supposed to affect the public reception of any movie, for example, such variables like movie/dvd release day/month/year are not meaningfull in the context of research question.

The variables mentioned above will be included in the full model. Then the model selection will be based on backward elimination using adjusted R squared, because the interest lies in seeking reliable prediction model, but not statistically significant predictors.

It means that one variable will be dropped while adjusted R-squared increases. The following code does what was described and prints details of the steps:

```
tot_vars<- c("genre", "big_studio", "runtime", "mpaa_rating",
"best_actor_win", "best_actress_win", "best_dir_win")

full_mod<- lm(avg_score ~ ., data = required_movies[c("avg_score",tot_vars)])

adj.r.sq.full <- summary(full_mod)$adj.r.squared

max.r.sq <- adj.r.sq.full

incl.vars <- tot_vars

repeat{
  n_vars <- length(incl.vars)
  drop.idx <- 0
  for(i in seq(n_vars)){
    mod <- lm(avg_score ~ ., data=required_movies[c("avg_score", incl.vars[-
i])])
    adj.r.sq <- summary(mod)$adj.r.squared
    cat("Drop:", incl.vars[i], "; Include:", incl.vars[-i], "\n")
    cat("Adj.R-squared:", round(adj.r.sq, 4), "\n")
    cat("*********", "\n")
    if(adj.r.sq > max.r.sq){
      max.r.sq <- adj.r.sq
      drop.idx <- i
    }
  }
  if(drop.idx != 0){
    incl.vars <- incl.vars[-drop.idx ]
  } else {
    break
  }
}
```

```
## Drop: genre ; Include: big_studio runtime mpaa_rating best_actor_win
best_actress_win best_dir_win
## Adj.R-squared: 0.1533
## *********
## Drop: big_studio ; Include: genre runtime mpaa_rating best_actor_win
best_actress_win best_dir_win
## Adj.R-squared: 0.2131
## *********
## Drop: runtime ; Include: genre big_studio mpaa_rating best_actor_win
best_actress_win best_dir_win
## Adj.R-squared: 0.1841
## *********
## Drop: mpaa_rating ; Include: genre big_studio runtime best_actor_win
best_actress_win best_dir_win
```

```
## Adj.R-squared: 0.1819
## *********
## Drop: best_actor_win ; Include: genre big_studio runtime mpaa_rating
best_actress_win best_dir_win
## Adj.R-squared: 0.2201
## *********
## Drop: best_actress_win ; Include: genre big_studio runtime mpaa_rating
best_actor_win best_dir_win
## Adj.R-squared: 0.2199
## *********
## Drop: best_dir_win ; Include: genre big_studio runtime mpaa_rating
best_actor_win best_actress_win
## Adj.R-squared: 0.2088
## *********
## Drop: genre ; Include: big_studio runtime mpaa_rating best_actress_win
best_dir_win
## Adj.R-squared: 0.1547
## *********
## Drop: big_studio ; Include: genre runtime mpaa_rating best_actress_win
best_dir_win
## Adj.R-squared: 0.2141
## *********
## Drop: runtime ; Include: genre big_studio mpaa_rating best_actress_win
best_dir_win
## Adj.R-squared: 0.1845
## *********
## Drop: mpaa_rating ; Include: genre big_studio runtime best_actress_win
best_dir_win
## Adj.R-squared: 0.1832
## *********
## Drop: best_actress_win ; Include: genre big_studio runtime mpaa_rating
best_dir_win
## Adj.R-squared: 0.2211
## *********
## Drop: best_dir_win ; Include: genre big_studio runtime mpaa_rating
best_actress_win
## Adj.R-squared: 0.21
## *********
## Drop: genre ; Include: big_studio runtime mpaa_rating best_dir_win
## Adj.R-squared: 0.1544
## *********
## Drop: big_studio ; Include: genre runtime mpaa_rating best_dir_win
## Adj.R-squared: 0.2153
## *********
## Drop: runtime ; Include: genre big_studio mpaa_rating best_dir_win
## Adj.R-squared: 0.1828
## *********
## Drop: mpaa_rating ; Include: genre big_studio runtime best_dir_win
## Adj.R-squared: 0.1845
## *********
```

```
## Drop: best_dir_win ; Include: genre big_studio runtime mpaa_rating
## Adj.R-squared: 0.2109
## *********
```

Hence now we come to the best model for the criteria as

```
best_mod <- lm(avg_score ~ ., data=required_movies[c("avg_score",
incl.vars)])
summary(best_mod)
```

```
##
## Call:
## lm(formula = avg_score ~ ., data = required_movies[c("avg_score",
##     incl.vars)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.092 -11.211   0.821  11.648  38.538
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    43.95520    6.13719   7.162 2.46e-12 ***
## genreAnimation                  0.07045    6.21332   0.011  0.99096
## genreArt House & International  5.97366    4.81650   1.240  0.21539
## genreComedy                     1.43743    2.61819   0.549  0.58321
## genreDrama                     11.66082    2.21517   5.264 2.00e-07 ***
## genreHorror                    -1.30001    3.91674  -0.332  0.74008
## genreMusical & Performing Arts 18.10078    5.89189   3.072  0.00223 **
## genreMystery & Suspense         5.53947    2.89083   1.916  0.05584 .
## genreOther                     13.95979    4.52351   3.086  0.00213 **
## genreScience Fiction & Fantasy -0.36390    5.56361  -0.065  0.94787
## big_studioTRUE                 -3.16305    1.38142  -2.290  0.02240 *
## runtime                         0.22431    0.04160   5.392 1.02e-07 ***
## mpaa_ratingNC-17               -6.88068   12.02241  -0.572  0.56733
## mpaa_ratingPG                 -12.77949    4.73496  -2.699  0.00716 **
## mpaa_ratingPG-13              -19.82451    4.81620  -4.116 4.42e-05 ***
## mpaa_ratingR                  -13.60985    4.70792  -2.891  0.00399 **
## mpaa_ratingUnrated             -4.32593    6.32131  -0.684  0.49404
## best_dir_winyes                 7.52845    2.58942   2.907  0.00379 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.61 on 570 degrees of freedom
## Multiple R-squared:  0.2436, Adjusted R-squared:  0.2211
## F-statistic:  10.8 on 17 and 570 DF,  p-value: < 2.2e-16
```

So the predictors to be included in the final model are : “genre”, “big studio”, “runtime”, “mpaa rating”, “best dir win”.

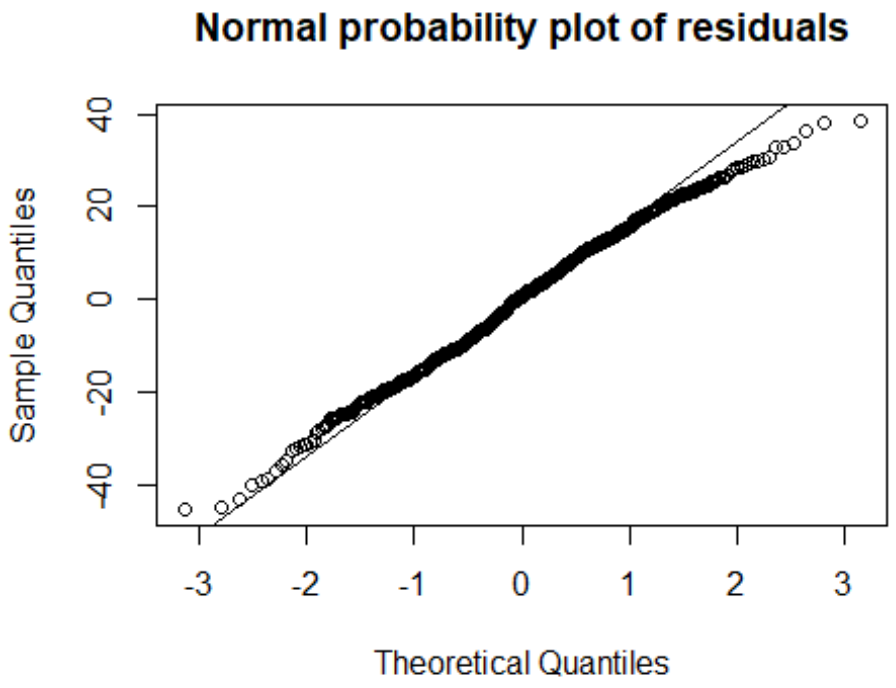
Best adjusted r squared value is : 0.2210541.

Also it's the model with all significant predictors. Particularly, it can be seen that while everything else held constant, movies of genre 'Drama' have 11.66 pts higher average score on average, 'Musical & Performing Arts' movies have 18.10 pts higher average score on average, movies produced by "big studios" have 3.16 pts less average score on average, etc. For example, movies with MPAA rating PG-13 have 19.82 pts less in their average score on average and movies with director who won Oscar have 7.52 pts higher score on average while everything else held constant.

Model Diagnostics:

The residuals are nearly normal with mean 0 :

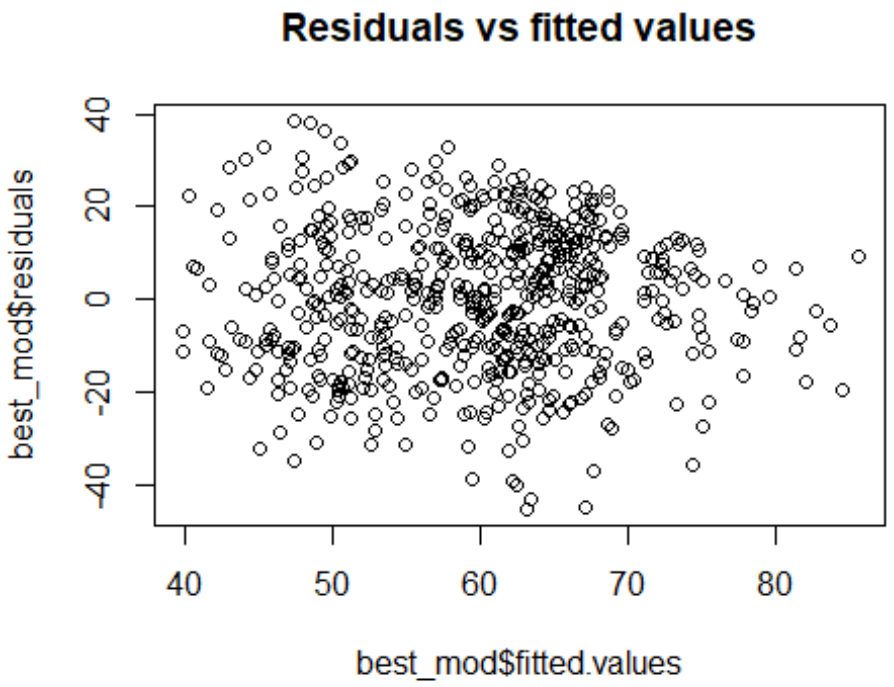
```
qqnorm(best_mod$residuals, main = "Normal probability plot of residuals")
qqline(best_mod$residuals)
```



As per the above plot it's true.

2.The variability of the residuals is nearly constant:

```
plot(best_mod$residuals ~ best_mod$fitted.values, main = "Residuals vs fitted
values")
```



Since there are no patterns, residuals are nearly randomly scattered.

The residuals are independent (observations are independent):

```
plot(best_mod$residuals, main = "Residuals in order of data collection")
```

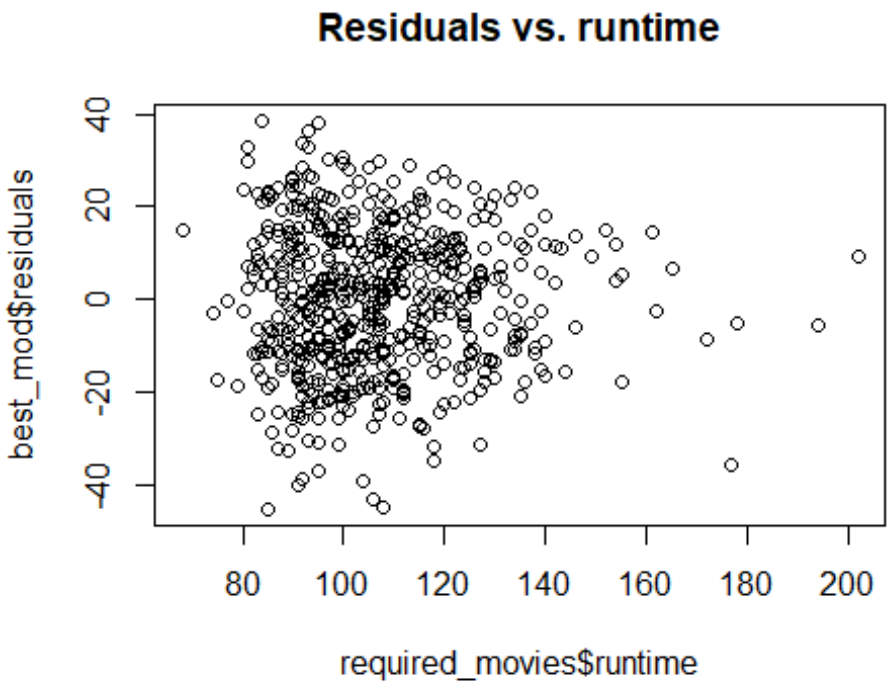
data collection above doesn't show any patterns.

It's known that data was collected using random sampling so this assumption also holds.

The linear relationship between the outcome and predictors.

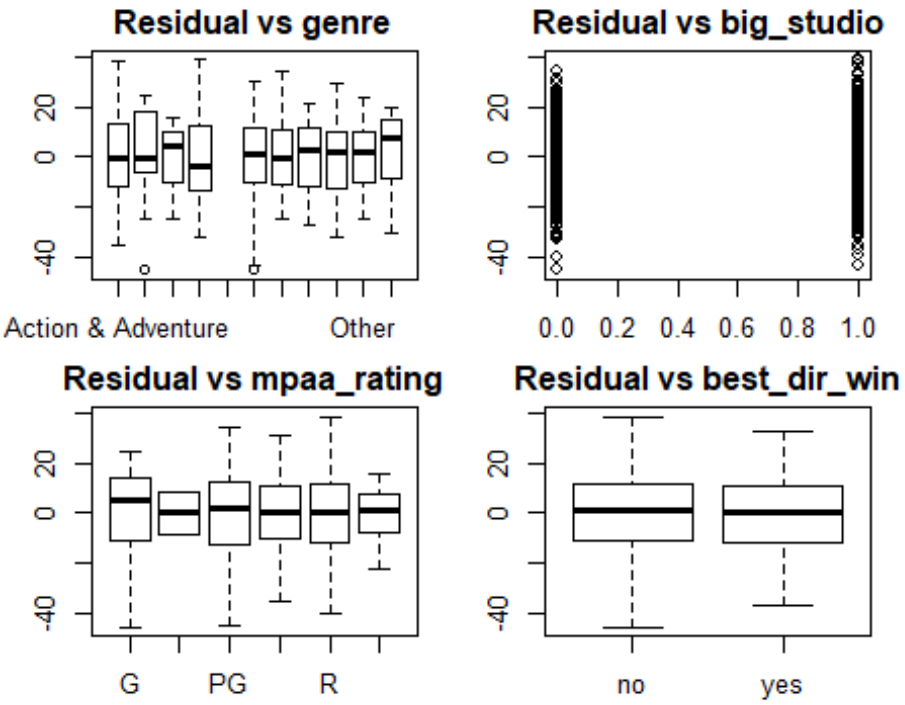
All predictors in the final model are categorical variables, except of runtime:

```
plot(best_mod$residuals ~ required_movies$runtime, main = "Residuals vs.
runtime")
```



The plot of the residuals versus runtime shows that residuals nearly randomly scattered around the 0. Also the plot of residuals versus each categorical variable shows nearly constant variabilty of residuals around 0 for each group:

```
par(mfrow=c(2,2), oma = c(1,1,1,1), mar = c(2,2,2,2))
plot(best_mod$residuals ~ required_movies$genre, main="Residual vs genre")
plot(best_mod$residuals ~ required_movies$big_studio, main="Residual vs
big_studio")
plot(best_mod$residuals ~ required_movies$mpaa_rating, main="Residual vs
mpaa_rating")
plot(best_mod$residuals ~ required_movies$best_dir_win, main="Residual vs
best_dir_win")
```



So it seems there is a relationship between the outcome and predictors is linear.

Part 5: Prediction

For the prediction part (2016) movie was selected, the source links for the movie: IMDB, Rotten Tomatoes.

```
movie = data.frame(genre="Comedy", runtime=106, mpaa_rating="PG-13",
best_dir_win="yes", big_studio=TRUE)
prediction <- predict(best_mod, movie, interval="predict", level = 0.95)
prediction
```

```
##        fit      lwr      upr
## 1 53.71074 22.42856 84.99292
```

The true avg score value for the movie is 64.66667 (the average of 3 scores: IMDB score 6.4 multiplied by 10, RT audience score 45%, RT critics score 85%). The 95% prediction interval for avg score predicted by final model is between 22.4285562 and 84.992921. The interval captures true value.

Part 6: Conclusion

This project determines the association between movie attributes and movie public reception resulting in multiple linear regression model. The association between movie genre, runtime, MPAA rating, production studio which produced many movies, director

who won Oscar and average score from IMDB audience and Rotten Tomatoes audience and critics were discovered.

Two shortcomings:

1.Intuitively public reception of movie is not just question of quantitative variables like the one presented in the dataset. In my opinion the public reception depends also on such creative things like novel idea, plot etc that hardly can be measured.

2.The model selection method was based on adjusted R-squared because the model was aimed at prediction, but not getting statistically significant predictors (which may result having diffent coefficients).