# Modeling and prediction for movies

Setup

Load packages

```
library(ggplot2)
library(dplyr)
library(statsr)
library(rio)
library(GGally)
library(tidyverse)
```

Load data

```
#load("movies.rda")
movies<-import("movies.rda")
```

---

Part 1: Data

The data set comprises 651 movies produced and released before 2016. The samples were randomly collected. The summary of the data (see below) also shows the sample has an adequate mix of category and value in responses. The large size and independence of the sample make it suitable for generalization of the population.

```
movies%>%
select(title_type,genre,runtime,mpaa_rating,thtr_rel_year:top200_box)%
>%summary()
```

```
##         title_type                  genre        runtime
mpaa_rating
##  Documentary : 55   Drama             :305   Min.   : 39.0   G
: 19
##  Feature Film:591   Comedy            : 87   1st Qu.: 92.0   NC-17
:  2
##  TV Movie    :  5   Action & Adventure: 65   Median :103.0   PG
:118
##                     Mystery & Suspense: 59   Mean   :105.8   PG-13
:133
##                     Documentary       : 52   3rd Qu.:115.8   R
:329
##                     Horror            : 23   Max.   :267.0
Unrated: 50
##                     (Other)           : 60   NA's   :1
##  thtr_rel_year  thtr_rel_month   thtr_rel_day    dvd_rel_year
dvd_rel_month
##  Min.   :1970   Min.   : 1.00   Min.   : 1.00   Min.   :1991   Min.
: 1.000
##  1st Qu.:1990   1st Qu.: 4.00   1st Qu.: 7.00   1st Qu.:2001   1st
Qu.: 3.000
##  Median :2000   Median : 7.00   Median :15.00   Median :2004
Median : 6.000
##  Mean   :1998   Mean   : 6.74   Mean   :14.42   Mean   :2004   Mean
: 6.333
##  3rd Qu.:2007   3rd Qu.:10.00   3rd Qu.:21.00   3rd Qu.:2008   3rd
Qu.: 9.000
##  Max.   :2014   Max.   :12.00   Max.   :31.00   Max.   :2015   Max.
:12.000
##                                                 NA's   :8      NA's
:8
##   dvd_rel_day     imdb_rating    imdb_num_votes
critics_rating
##  Min.   : 1.00   Min.   :1.900   Min.   :   180   Certified
Fresh:135
##  1st Qu.: 7.00   1st Qu.:5.900   1st Qu.:  4546   Fresh
:209
##  Median :15.00   Median :6.600   Median : 15116   Rotten
:307
##  Mean   :15.01   Mean   :6.493   Mean   : 57533
##  3rd Qu.:23.00   3rd Qu.:7.300   3rd Qu.: 58301
##  Max.   :31.00   Max.   :9.000   Max.   :893008
##  NA's   :8
##  critics_score    audience_rating audience_score  best_pic_nom
best_pic_win
##  Min.   :  1.00   Spilled:275     Min.   :11.00   no :629      no
:644
##  1st Qu.: 33.00   Upright:376     1st Qu.:46.00   yes: 22      yes:
```

```
7
##  Median : 61.00                    Median :65.00
##  Mean   : 57.69                    Mean   :62.36
##  3rd Qu.: 83.00                    3rd Qu.:80.00
##  Max.   :100.00                    Max.   :97.00
##
##  best_actor_win best_actress_win best_dir_win top200_box
##  no :558        no :579          no :608      no :636
##  yes: 93        yes: 72          yes: 43      yes: 15
##
##
##
##
##
```

---

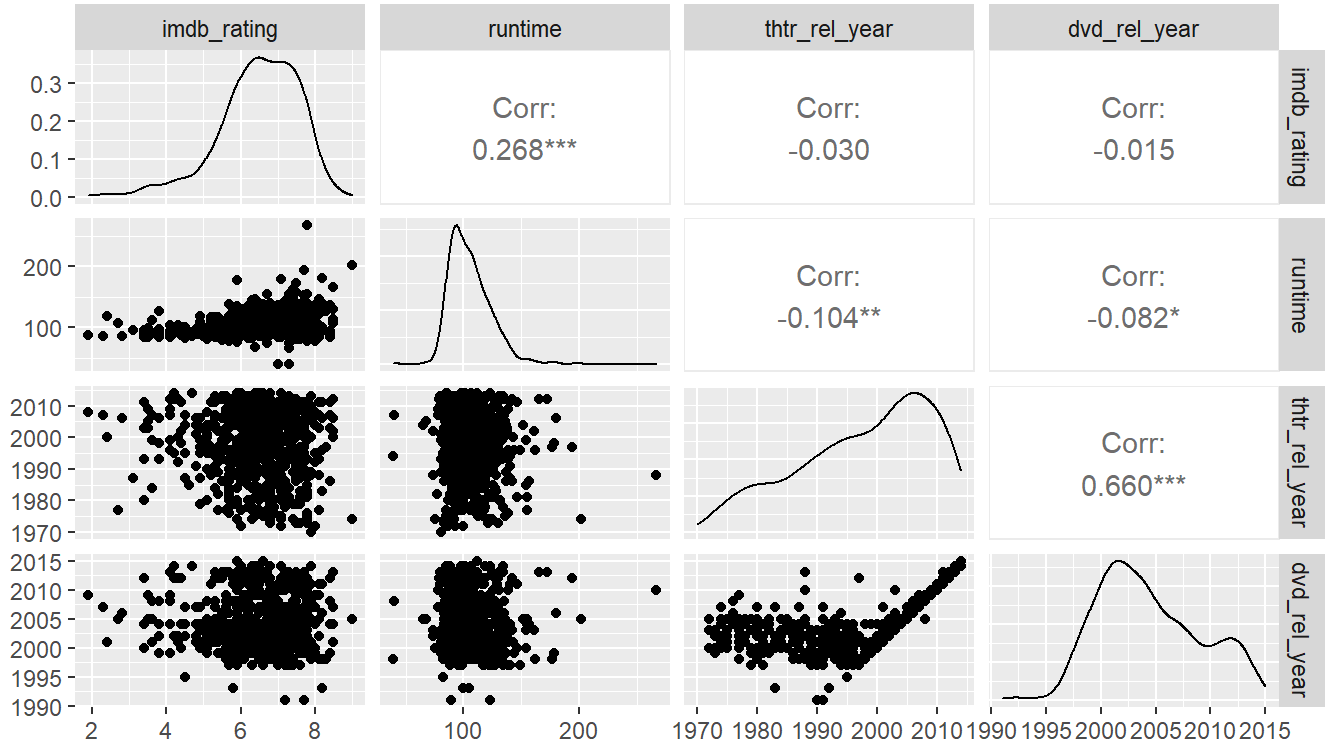
Part 2: Research question

Research question: “How to predict imbd rating of a movies before it is released?”
Why should  this question be in the interest of the boss? Because imdb rating is an important factor that shows the success of a movie, as well as its popularity. Predicting imdb rating before it is released therefore can be beneficial for commercial purposes.

---

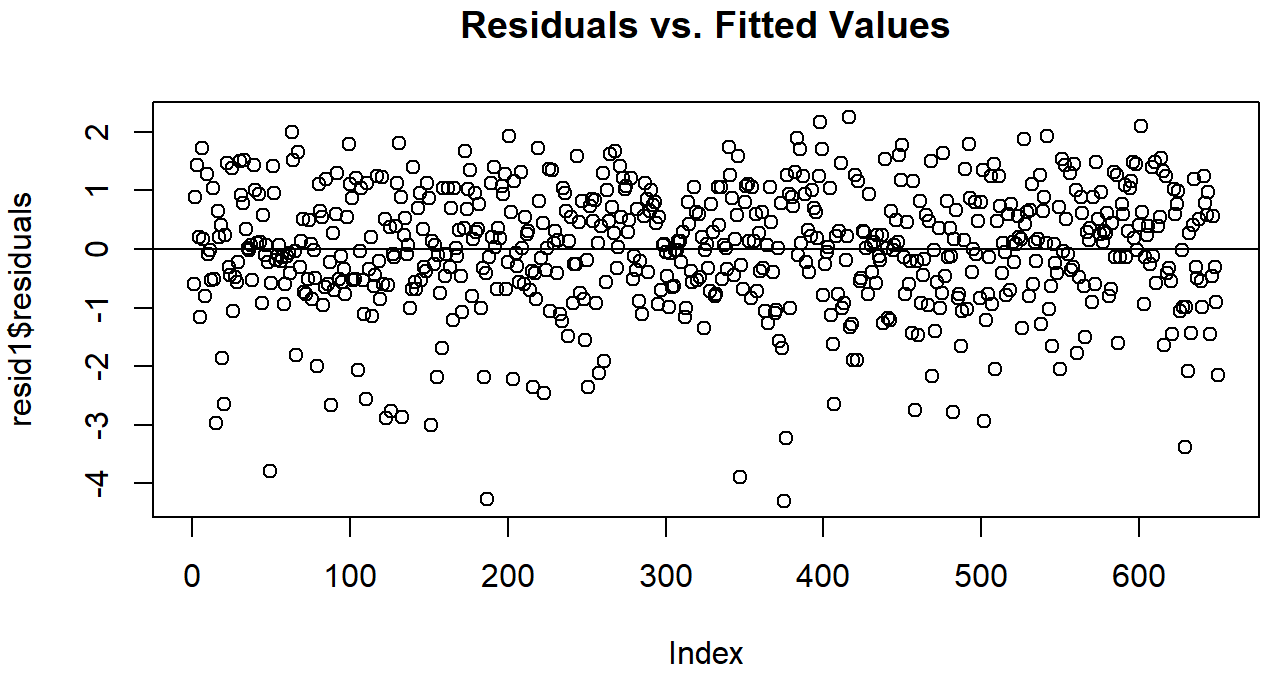
Part 3: Exploratory data analysis

```
movies %>% select(imdb_rating, runtime, thtr_rel_year,
dvd_rel_year)%>%
     ggpairs()
```

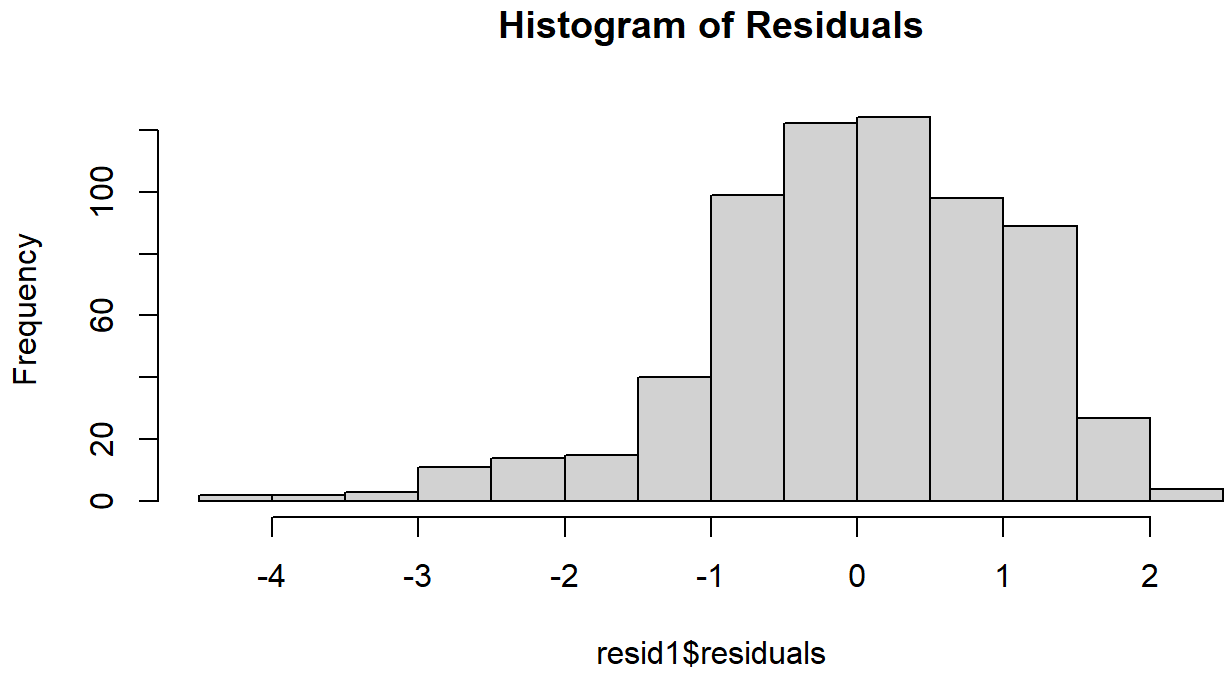



This diagram shows the correlation coefficients amongst four numerical variables in the chosen multiple regression model. The result indicates very weak correlations between imdb rating and thr_rel_year, and dvd_rel_year, at -0.03 and -0.015, while the figure for correlation between imbd rating and runtime is more significant, at 0.268.
To check if the conditions of least squares regression (between imbd rating and runtime) are met, residual plots (scatter plot of residuals, histogram & probability plot of residuals) are created as the following:

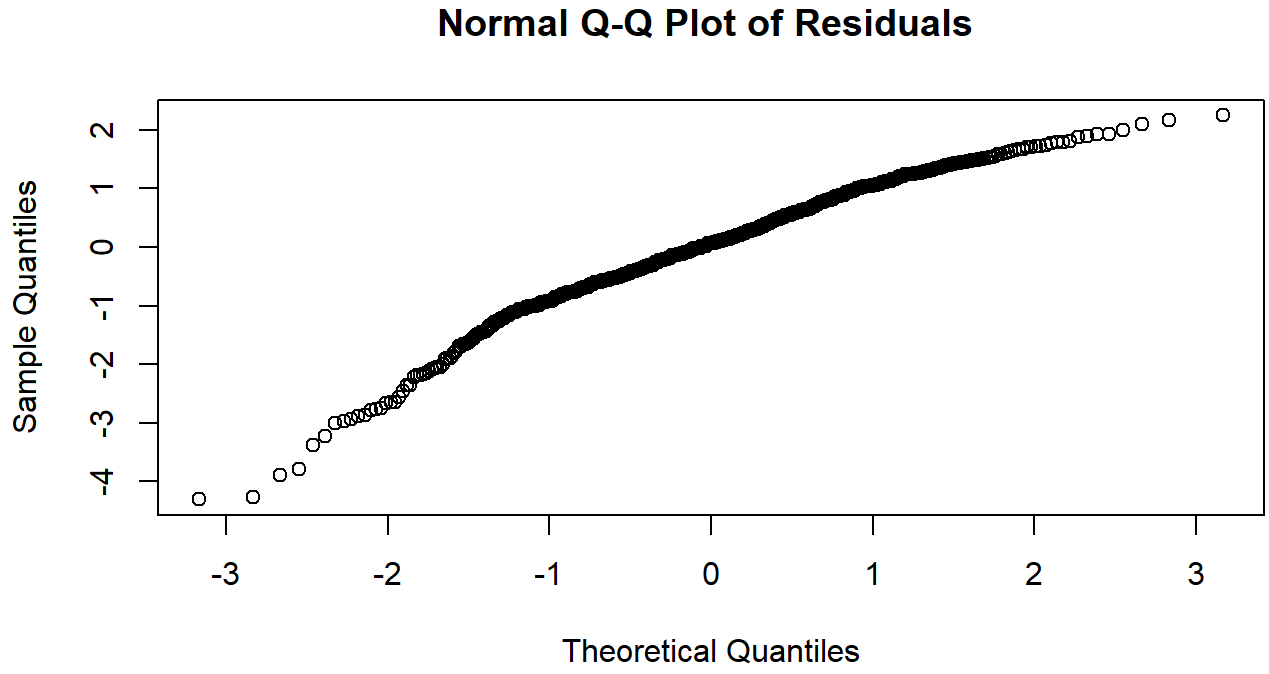
```
resid1<- lm(imdb_rating~runtime, data=movies)

plot(resid1$residuals, main= "Residuals vs. Fitted Values")
abline(h=0)
```




```
hist(resid1$residuals, main= "Histogram of Residuals")
```




```
qqnorm(resid1$residuals, main= "Normal Q-Q Plot of Residuals ")
```



The residual plot shows a random scatter around the zero line, no fan shape, indicating a linear relationship between variables and constant variation of residuals. The histogram and Q-Q plot also show a normal distribution of residuals.

---

Part 4: Modeling

As the model is to predict the imdb rating of a movie before it is released, the variables must be elements that are known before the release date. Hence, variables being excluded are imdb_num_vote, critics_rating, critics_score, audience_rating, audience_score, best_pic_nom, best_pic_win, top 200_box. Next, variable regarding title of movie, day and month, actor/actress name are unuseful for regression model only for information purpose as the wide range of category variables are not suitable for the modeling. Two variables thtr_rel_year and dvd_rel_year are also omitted due to extremely weak linear correlation with imdb rating as demonstrated in Part 3. Thus, the full model starts off with response variables: runtime, title_type, genre, mpaa_rating, best_dir_win, best_actress_win, best_actor_win.

Clean data

```
#checking NA row
movies%>% select(imdb_rating, runtime, title_type, genre, mpaa_rating,
best_dir_win, best_actress_win, best_actor_win) %>% summary()
##   imdb_rating       runtime             title_type
genre
##  Min.   :1.900   Min.   : 39.0   Documentary : 55   Drama
:305
##  1st Qu.:5.900   1st Qu.: 92.0   Feature Film:591   Comedy
: 87
##  Median :6.600   Median :103.0   TV Movie    :  5   Action &
Adventure: 65
##  Mean   :6.493   Mean   :105.8                      Mystery &
Suspense: 59
##  3rd Qu.:7.300   3rd Qu.:115.8                      Documentary
: 52
##  Max.   :9.000   Max.   :267.0                      Horror
: 23
##                  NA's   :1                          (Other)
: 60
##   mpaa_rating  best_dir_win best_actress_win best_actor_win
##  G      : 19   no :608      no :579          no :558
##  NC-17  :  2   yes: 43      yes: 72          yes: 93
##  PG     :118
##  PG-13  :133
```

```
##  R      :329
##  Unrated: 50
##
```

```
#removing NA row
movies1<- movies %>% select(imdb_rating, runtime, title_type, genre,
mpaa_rating, best_dir_win, best_actress_win, best_actor_win) %>%
na.omit()
```

Split data

```
set.seed(2)
train <-movies1%>% sample_frac(.70)
test <-anti_join(movies1,train)
## Joining, by = c("imdb_rating", "runtime", "title_type", "genre",
"mpaa_rating", "best_dir_win", "best_actress_win", "best_actor_win")
```

Model selection

As the aim to build the model is for prediction. The method for model selection is to use adjusted R_square, and with backward elimination.

```
#Full model
fit=
lm(imdb_rating~runtime+title_type+genre+mpaa_rating+best_dir_win+best_
actress_win+ best_actor_win, data=train)

#Remove one variable at a time
fit1=
lm(imdb_rating~runtime+title_type+genre+mpaa_rating+best_dir_win+best_
actress_win, data=train)
fit2=
lm(imdb_rating~runtime+title_type+genre+mpaa_rating+best_dir_win+
best_actor_win, data=train)
fit3=
lm(imdb_rating~runtime+title_type+genre+mpaa_rating+best_actress_win+
best_actor_win, data=train)
fit4=
```

```
lm(imdb_rating~runtime+title_type+genre+best_dir_win+best_actress_win+
best_actor_win, data=train)
fit5=
lm(imdb_rating~runtime+title_type+mpaa_rating+best_dir_win+best_actres
s_win+ best_actor_win, data=train)
fit6=
lm(imdb_rating~runtime+genre+mpaa_rating+best_dir_win+best_actress_win
+ best_actor_win, data=train)
summary(fit)
##
## Call:
## lm(formula = imdb_rating ~ runtime + title_type + genre +
mpaa_rating +
##     best_dir_win + best_actress_win + best_actor_win, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7968 -0.5194  0.0952  0.5694  1.9497
##
## Coefficients:
##                                 Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                     6.874761   0.666130  10.320  <
2e-16 ***
## runtime                         0.012043   0.002391   5.036
6.99e-07 ***
## title_typeFeature Film         -1.127171   0.516943  -2.180
0.029761 *
## title_typeTV Movie             -2.127945   0.870165  -2.445
0.014864 *
## genreAnimation                 -0.884210   0.504704  -1.752
0.080492 .
## genreArt House & International  0.340942   0.332558   1.025
0.305837
## genreComedy                    -0.158471   0.190116  -0.834
0.404996
## genreDocumentary                0.520967   0.564229   0.923
0.356352
## genreDrama                      0.573889   0.159810   3.591
0.000367 ***
## genreHorror                    -0.079836   0.268078  -0.298
0.765992
## genreMusical & Performing Arts  0.286702   0.445037   0.644
0.519773
## genreMystery & Suspense         0.267992   0.209210   1.281
0.200888
## genreOther                      0.626712   0.331432   1.891
0.059302 .
```

```
## genreScience Fiction & Fantasy -0.569808   0.381535  -1.493
0.136045
## mpaa_ratingNC-17               -0.884727   0.748039  -1.183
0.237566
## mpaa_ratingPG                  -1.101247   0.353087  -3.119
0.001936 **
## mpaa_ratingPG-13               -1.266361   0.356519  -3.552
0.000424 ***
## mpaa_ratingR                   -0.943074   0.349879  -2.695
0.007304 **
## mpaa_ratingUnrated             -0.874988   0.407302  -2.148
0.032248 *
## best_dir_winyes                 0.527309   0.174734   3.018
0.002697 **
## best_actress_winyes             0.124047   0.145815   0.851
0.395397
## best_actor_winyes               0.117891   0.135194   0.872
0.383686
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9323 on 433 degrees of freedom
## Multiple R-squared:  0.3223, Adjusted R-squared:  0.2895
## F-statistic: 9.808 on 21 and 433 DF,  p-value: < 2.2e-16
```

```
summary(fit1)
##
## Call:
## lm(formula = imdb_rating ~ runtime + title_type + genre +
mpaa_rating +
##     best_dir_win + best_actress_win, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8185 -0.5311  0.1045  0.5567  1.9371
##
## Coefficients:
##                                Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                    6.813610   0.662246  10.289  <
2e-16 ***
## runtime                        0.012520   0.002327   5.380
1.22e-07 ***
## title_typeFeature Film        -1.108514   0.516358  -2.147
0.032363 *
```

```
## title_typeTV Movie             -2.127622   0.869925  -2.446
0.014851 *
## genreAnimation                 -0.861240   0.503877  -1.709
0.088124 .
## genreArt House & International  0.328818   0.332176   0.990
0.322778
## genreComedy                    -0.155314   0.190029  -0.817
0.414196
## genreDocumentary                0.531370   0.563947   0.942
0.346597
## genreDrama                      0.576731   0.159733   3.611
0.000341 ***
## genreHorror                    -0.086586   0.267893  -0.323
0.746692
## genreMusical & Performing Arts  0.292885   0.444858   0.658
0.510644
## genreMystery & Suspense         0.277731   0.208854   1.330
0.184287
## genreOther                      0.619474   0.331236   1.870
0.062131 .
## genreScience Fiction & Fantasy -0.583075   0.381126  -1.530
0.126777
## mpaa_ratingNC-17               -0.833857   0.745555  -1.118
0.263999
## mpaa_ratingPG                  -1.090488   0.352774  -3.091
0.002122 **
## mpaa_ratingPG-13               -1.265767   0.356420  -3.551
0.000425 ***
## mpaa_ratingR                   -0.938868   0.349749  -2.684
0.007544 **
## mpaa_ratingUnrated             -0.870997   0.407164  -2.139
0.032978 *
## best_dir_winyes                 0.522313   0.174592   2.992
0.002933 **
## best_actress_winyes             0.137538   0.144952   0.949
0.343224
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.932 on 434 degrees of freedom
## Multiple R-squared:  0.3211, Adjusted R-squared:  0.2899
## F-statistic: 10.27 on 20 and 434 DF,  p-value: < 2.2e-16



summary(fit2)
##
```

```
## Call:
## lm(formula = imdb_rating ~ runtime + title_type + genre +
mpaa_rating +
##     best_dir_win + best_actor_win, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8153 -0.5236  0.0912  0.5742  1.9416
##
## Coefficients:
##                                 Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                     6.848981   0.665229  10.296  <
2e-16 ***
## runtime                         0.012374   0.002359   5.246
2.44e-07 ***
## title_typeFeature Film         -1.123059   0.516756  -2.173
0.030299 *
## title_typeTV Movie             -2.067532   0.866986  -2.385
0.017520 *
## genreAnimation                 -0.891123   0.504478  -1.766
0.078027 .
## genreArt House & International  0.358034   0.331845   1.079
0.281224
## genreComedy                    -0.137236   0.188411  -0.728
0.466771
## genreDocumentary                0.535186   0.563802   0.949
0.343025
## genreDrama                      0.590320   0.158588   3.722
0.000223 ***
## genreHorror                    -0.072057   0.267837  -0.269
0.788033
## genreMusical & Performing Arts  0.283766   0.444882   0.638
0.523910
## genreMystery & Suspense         0.280721   0.208608   1.346
0.179106
## genreOther                      0.637288   0.331093   1.925
0.054908 .
## genreScience Fiction & Fantasy -0.569009   0.381413  -1.492
0.136466
## mpaa_ratingNC-17               -0.918714   0.746734  -1.230
0.219248
## mpaa_ratingPG                  -1.112645   0.352720  -3.154
0.001720 **
## mpaa_ratingPG-13               -1.282235   0.355917  -3.603
0.000351 ***
## mpaa_ratingR                   -0.960028   0.349200  -2.749
0.006223 **
```

```
## mpaa_ratingUnrated             -0.896820   0.406363  -2.207
0.027841 *
## best_dir_winyes                 0.531532   0.174608   3.044
0.002475 **
## best_actor_winyes               0.130093   0.134388   0.968
0.333564
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.932 on 434 degrees of freedom
## Multiple R-squared:  0.3212, Adjusted R-squared:  0.2899
## F-statistic: 10.27 on 20 and 434 DF,  p-value: < 2.2e-16
```

```
summary(fit3)
##
## Call:
## lm(formula = imdb_rating ~ runtime + title_type + genre +
mpaa_rating +
##     best_actress_win + best_actor_win, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8514 -0.5079  0.0747  0.5864  2.3118
##
## Coefficients:
##                                 Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                     6.631785   0.667394   9.937  <
2e-16 ***
## runtime                         0.013849   0.002337   5.926
6.32e-09 ***
## title_typeFeature Film         -1.048983   0.521093  -2.013
0.044727 *
## title_typeTV Movie             -2.050373   0.877871  -2.336
0.019966 *
## genreAnimation                 -0.878622   0.509392  -1.725
0.085268 .
## genreArt House & International  0.300766   0.335381   0.897
0.370329
## genreComedy                    -0.152536   0.191874  -0.795
0.427060
## genreDocumentary                0.577936   0.569156   1.015
0.310468
## genreDrama                      0.567204   0.161280   3.517
0.000483 ***
```

```
## genreHorror                    -0.103089   0.270459  -0.381
0.703267
## genreMusical & Performing Arts  0.330209   0.448939   0.736
0.462412
## genreMystery & Suspense         0.277193   0.211132   1.313
0.189915
## genreOther                      0.568325   0.333943   1.702
0.089497 .
## genreScience Fiction & Fantasy -0.542812   0.384976  -1.410
0.159259
## mpaa_ratingNC-17               -0.887156   0.754993  -1.175
0.240618
## mpaa_ratingPG                  -1.066609   0.356181  -2.995
0.002906 **
## mpaa_ratingPG-13               -1.271127   0.359830  -3.533
0.000456 ***
## mpaa_ratingR                   -0.920084   0.353048  -2.606
0.009473 **
## mpaa_ratingUnrated             -0.887934   0.411066  -2.160
0.031313 *
## best_actress_winyes             0.136546   0.147111   0.928
0.353827
## best_actor_winyes               0.104512   0.136378   0.766
0.443887
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.941 on 434 degrees of freedom
## Multiple R-squared:  0.3081, Adjusted R-squared:  0.2762
## F-statistic: 9.662 on 20 and 434 DF,  p-value: < 2.2e-16
```

```
summary(fit4)
##
## Call:
## lm(formula = imdb_rating ~ runtime + title_type + genre +
best_dir_win +
##     best_actress_win + best_actor_win, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8674 -0.5336  0.0860  0.5894  2.1417
##
## Coefficients:
##                                 Estimate Std. Error t value
Pr(>|t|)
```

```
## (Intercept)                      6.041408   0.604909   9.987  <
2e-16 ***
## runtime                          0.011274   0.002387   4.723
3.13e-06 ***
## title_typeFeature Film          -1.197764   0.523829  -2.287
0.022698 *
## title_typeTV Movie              -2.053406   0.855081  -2.401
0.016747 *
## genreAnimation                  -0.134923   0.448625  -0.301
0.763750
## genreArt House & International   0.354918   0.332311   1.068
0.286096
## genreComedy                     -0.267847   0.189715  -1.412
0.158708
## genreDocumentary                 0.520792   0.553873   0.940
0.347596
## genreDrama                       0.528716   0.158368   3.339
0.000914 ***
## genreHorror                     -0.057563   0.265739  -0.217
0.828610
## genreMusical & Performing Arts   0.298918   0.448719   0.666
0.505660
## genreMystery & Suspense          0.236857   0.207264   1.143
0.253755
## genreOther                       0.567259   0.332653   1.705
0.088855 .
## genreScience Fiction & Fantasy -0.644827   0.385392  -1.673
0.095008 .
## best_dir_winyes                  0.544257   0.176242   3.088
0.002142 **
## best_actress_winyes              0.134264   0.147386   0.911
0.362810
## best_actor_winyes                0.119742   0.136260   0.879
0.380005
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.946 on 438 degrees of freedom
## Multiple R-squared:  0.2942, Adjusted R-squared:  0.2684
## F-statistic: 11.41 on 16 and 438 DF,  p-value: < 2.2e-16
```

```
summary(fit5)
##
## Call:
## lm(formula = imdb_rating ~ runtime + title_type + mpaa_rating +
```

```
##     best_dir_win + best_actress_win + best_actor_win, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8525 -0.4938  0.1212  0.6309  2.0794
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               6.530949   0.420621  15.527  < 2e-16 ***
## runtime                   0.015537   0.002403   6.467 2.65e-10 ***
## title_typeFeature Film -1.295849   0.236310  -5.484 7.01e-08 ***
## title_typeTV Movie     -2.081734   0.720793  -2.888  0.00407 **
## mpaa_ratingNC-17       -0.152972   0.751042  -0.204  0.83870
## mpaa_ratingPG          -0.695931   0.313686  -2.219  0.02702 *
## mpaa_ratingPG-13       -0.834254   0.313560  -2.661  0.00808 **
## mpaa_ratingR           -0.467224   0.303005  -1.542  0.12380
## mpaa_ratingUnrated     -0.378402   0.372409  -1.016  0.31014
## best_dir_winyes         0.480012   0.181495   2.645  0.00847 **
## best_actress_winyes     0.182129   0.150225   1.212  0.22602
## best_actor_winyes       0.128425   0.140091   0.917  0.35979
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9736 on 443 degrees of freedom
## Multiple R-squared:  0.2438, Adjusted R-squared:  0.2251
## F-statistic: 12.99 on 11 and 443 DF,  p-value: < 2.2e-16
```

```
summary(fit6)
##
## Call:
## lm(formula = imdb_rating ~ runtime + genre + mpaa_rating +
best_dir_win +
##     best_actress_win + best_actor_win, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7887 -0.5083  0.0837  0.5661  1.9203
##
## Coefficients:
##                                 Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                     5.698323   0.410437  13.884  <
2e-16 ***
## runtime                         0.012575   0.002395   5.250
2.38e-07 ***
```

```
## genreAnimation                 -0.877787   0.507479  -1.730
0.084393 .
## genreArt House & International  0.346957   0.334284   1.038
0.299889
## genreComedy                    -0.133889   0.190903  -0.701
0.483461
## genreDocumentary                1.670990   0.274204   6.094
2.43e-09 ***
## genreDrama                      0.566204   0.160640   3.525
0.000469 ***
## genreHorror                    -0.069086   0.269374  -0.256
0.797711
## genreMusical & Performing Arts  0.668391   0.413942   1.615
0.107101
## genreMystery & Suspense         0.265155   0.210366   1.260
0.208183
## genreOther                      0.633552   0.333164   1.902
0.057881 .
## genreScience Fiction & Fantasy -0.569757   0.383653  -1.485
0.138246
## mpaa_ratingNC-17               -0.877739   0.752165  -1.167
0.243870
## mpaa_ratingPG                  -1.108814   0.355012  -3.123
0.001908 **
## mpaa_ratingPG-13               -1.272624   0.358434  -3.551
0.000426 ***
## mpaa_ratingR                   -0.942203   0.351784  -2.678
0.007679 **
## mpaa_ratingUnrated             -0.971735   0.402161  -2.416
0.016091 *
## best_dir_winyes                 0.508119   0.175483   2.896
0.003976 **
## best_actress_winyes             0.101321   0.145972   0.694
0.487982
## best_actor_winyes               0.111393   0.135770   0.820
0.412405
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9375 on 435 degrees of freedom
## Multiple R-squared:  0.3116, Adjusted R-squared:  0.2816
## F-statistic: 10.36 on 19 and 435 DF,  p-value: < 2.2e-16
```

Fit1 and fit2 models are selected due to highest adjusted R2, both at 0.2899. Hence, the next step is to try to remove both best_actress_win and best_actor_win.

```
#Remove one variable at a time
fit7=
lm(imdb_rating~runtime+title_type+genre+mpaa_rating+best_dir_win,
data=train)
summary(fit7)
##
## Call:
## lm(formula = imdb_rating ~ runtime + title_type + genre +
mpaa_rating +
##     best_dir_win, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8417 -0.5361  0.0922  0.5712  1.9267
##
## Coefficients:
##                                 Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                     6.777604   0.661082  10.252  <
2e-16 ***
## runtime                         0.012947   0.002283   5.671
2.6e-08 ***
## title_typeFeature Film         -1.101736   0.516249  -2.134
0.033392 *
## title_typeTV Movie             -2.059838   0.866887  -2.376
0.017927 *
## genreAnimation                 -0.866325   0.503790  -1.720
0.086214 .
## genreArt House & International  0.346577   0.331610   1.045
0.296541
## genreComedy                    -0.131135   0.188292  -0.696
0.486522
## genreDocumentary                0.548523   0.563593   0.973
0.330965
## genreDrama                      0.595487   0.158487   3.757
0.000195 ***
## genreHorror                    -0.078645   0.267731  -0.294
0.769091
## genreMusical & Performing Arts  0.290310   0.444798   0.653
0.514309
## genreMystery & Suspense         0.293135   0.208198   1.408
0.159857
## genreOther                      0.630494   0.330995   1.905
0.057460 .
## genreScience Fiction & Fantasy -0.583719   0.381082  -1.532
0.126313
## mpaa_ratingNC-17               -0.866065   0.744697  -1.163
0.245477
```

```
## mpaa_ratingPG                  -1.102021   0.352524  -3.126
0.001890 **
## mpaa_ratingPG-13               -1.283499   0.355889  -3.606
0.000346 ***
## mpaa_ratingR                   -0.957392   0.349164  -2.742
0.006359 **
## mpaa_ratingUnrated             -0.895016   0.406330  -2.203
0.028141 *
## best_dir_winyes                 0.526468   0.174517   3.017
0.002705 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9319 on 435 degrees of freedom
## Multiple R-squared:  0.3197, Adjusted R-squared:   0.29
## F-statistic: 10.76 on 19 and 435 DF,  p-value: < 2.2e-16
```

```
fit8= lm(imdb_rating~runtime+title_type+genre+mpaa_rating, data=train)
fit9= lm(imdb_rating~runtime+title_type+genre+best_dir_win,
data=train)
fit10= lm(imdb_rating~runtime+title_type+mpaa_rating+best_dir_win,
data=train)
fit11= lm(imdb_rating~runtime+genre+mpaa_rating+best_dir_win,
data=train)
summary(fit8)
##
## Call:
## lm(formula = imdb_rating ~ runtime + title_type + genre +
mpaa_rating,
##     data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8957 -0.5232  0.0732  0.5761  2.4094
##
## Coefficients:
##                                Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                    6.538671   0.662389   9.871  <
2e-16 ***
## runtime                        0.014733   0.002226   6.620
1.06e-10 ***
## title_typeFeature Film        -1.025117   0.520392  -1.970
0.049483 *
## title_typeTV Movie            -1.976888   0.874463  -2.261
```

```
0.024271 *
## genreAnimation                  -0.863749   0.508448  -1.699
0.090071 .
## genreArt House & International  0.309189   0.334443   0.924
0.355742
## genreComedy                     -0.123629   0.190016  -0.651
0.515633
## genreDocumentary                 0.605686   0.568483   1.065
0.287265
## genreDrama                       0.589974   0.159942   3.689
0.000254 ***
## genreHorror                     -0.100502   0.270108  -0.372
0.710013
## genreMusical & Performing Arts  0.332915   0.448685   0.742
0.458500
## genreMystery & Suspense          0.302459   0.210100   1.440
0.150699
## genreOther                       0.573801   0.333517   1.720
0.086059 .
## genreScience Fiction & Fantasy -0.555266   0.384488  -1.444
0.149409
## mpaa_ratingNC-17                -0.876837   0.751574  -1.167
0.243983
## mpaa_ratingPG                   -1.069520   0.355617  -3.008
0.002786 **
## mpaa_ratingPG-13                -1.289747   0.359174  -3.591
0.000367 ***
## mpaa_ratingR                    -0.936354   0.352323  -2.658
0.008158 **
## mpaa_ratingUnrated              -0.910333   0.410055  -2.220
0.026931 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9405 on 436 degrees of freedom
## Multiple R-squared:  0.3055, Adjusted R-squared:  0.2768
## F-statistic: 10.66 on 18 and 436 DF,  p-value: < 2.2e-16
```

```
summary(fit9)
##
## Call:
## lm(formula = imdb_rating ~ runtime + title_type + genre +
best_dir_win,
##     data = train)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9023 -0.5167  0.0733  0.5794  2.1420
##
## Coefficients:
##                                 Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                     5.932466   0.599146   9.902  <
2e-16 ***
## runtime                         0.012174   0.002287   5.322
1.64e-07 ***
## title_typeFeature Film         -1.170106   0.523183  -2.237
0.025818 *
## title_typeTV Movie             -1.987208   0.852332  -2.331
0.020178 *
## genreAnimation                 -0.106900   0.447856  -0.239
0.811454
## genreArt House & International  0.356400   0.331583   1.075
0.283034
## genreComedy                    -0.240704   0.188140  -1.279
0.201435
## genreDocumentary                0.545450   0.553438   0.986
0.324888
## genreDrama                      0.549502   0.157329   3.493
0.000527 ***
## genreHorror                    -0.059868   0.265527  -0.225
0.821719
## genreMusical & Performing Arts  0.300310   0.448500   0.670
0.503472
## genreMystery & Suspense         0.259572   0.206520   1.257
0.209462
## genreOther                      0.573876   0.332346   1.727
0.084915 .
## genreScience Fiction & Fantasy -0.657929   0.385070  -1.709
0.088231 .
## best_dir_winyes                 0.545346   0.176065   3.097
0.002077 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9458 on 440 degrees of freedom
## Multiple R-squared:  0.2913, Adjusted R-squared:  0.2688
## F-statistic: 12.92 on 14 and 440 DF,  p-value: < 2.2e-16
```

```
summary(fit10)
```

```
##
## Call:
## lm(formula = imdb_rating ~ runtime + title_type + mpaa_rating +
##     best_dir_win, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8646 -0.4983  0.1148  0.6247  2.0175
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              6.435319   0.416013  15.469  < 2e-16 ***
## runtime                  0.016695   0.002284   7.309 1.25e-12 ***
## title_typeFeature Film -1.276761   0.236081  -5.408 1.04e-07 ***
## title_typeTV Movie     -1.987677   0.717181  -2.772  0.00581 **
## mpaa_ratingNC-17       -0.128103   0.749450  -0.171  0.86436
## mpaa_ratingPG          -0.692490   0.313874  -2.206  0.02788 *
## mpaa_ratingPG-13       -0.846447   0.313657  -2.699  0.00723 **
## mpaa_ratingR           -0.478434   0.303126  -1.578  0.11520
## mpaa_ratingUnrated     -0.400492   0.372390  -1.075  0.28275
## best_dir_winyes         0.481497   0.181500   2.653  0.00827 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9743 on 445 degrees of freedom
## Multiple R-squared:  0.2394, Adjusted R-squared:  0.224
## F-statistic: 15.56 on 9 and 445 DF,  p-value: < 2.2e-16
```

```
summary(fit11)
##
## Call:
## lm(formula = imdb_rating ~ runtime + genre + mpaa_rating +
best_dir_win,
##     data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8289 -0.5286  0.0826  0.5771  1.9337
##
## Coefficients:
##                                Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                    5.635145   0.406102  13.876  <
2e-16 ***
## runtime                        0.013369   0.002288   5.843
```

```
1.00e-08 ***
## genreAnimation                  -0.860541   0.506328  -1.700
0.089923 .
## genreArt House & International  0.350039   0.333238   1.050
0.294108
## genreComedy                     -0.111271   0.189023  -0.589
0.556391
## genreDocumentary                 1.670367   0.273823   6.100
2.34e-09 ***
## genreDrama                       0.584780   0.159183   3.674
0.000269 ***
## genreHorror                     -0.069244   0.268945  -0.257
0.796941
## genreMusical & Performing Arts  0.664028   0.413561   1.606
0.109077
## genreMystery & Suspense         0.287304   0.209246   1.373
0.170442
## genreOther                       0.635616   0.332613   1.911
0.056661 .
## genreScience Fiction & Fantasy -0.582837   0.383033  -1.522
0.128824
## mpaa_ratingNC-17                -0.856784   0.748426  -1.145
0.252926
## mpaa_ratingPG                   -1.108143   0.354280  -3.128
0.001878 **
## mpaa_ratingPG-13                -1.286804   0.357628  -3.598
0.000357 ***
## mpaa_ratingR                    -0.953785   0.350887  -2.718
0.006825 **
## mpaa_ratingUnrated              -0.984608   0.401398  -2.453
0.014559 *
## best_dir_winyes                  0.507261   0.175197   2.895
0.003977 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9367 on 437 degrees of freedom
## Multiple R-squared:  0.3096, Adjusted R-squared:  0.2827
## F-statistic: 11.53 on 17 and 437 DF,  p-value: < 2.2e-16
```

Hence, fit7 model is selected for final model as none of later models yield an increase in adjusted R2.

```
summary(fit7)
##
## Call:
```

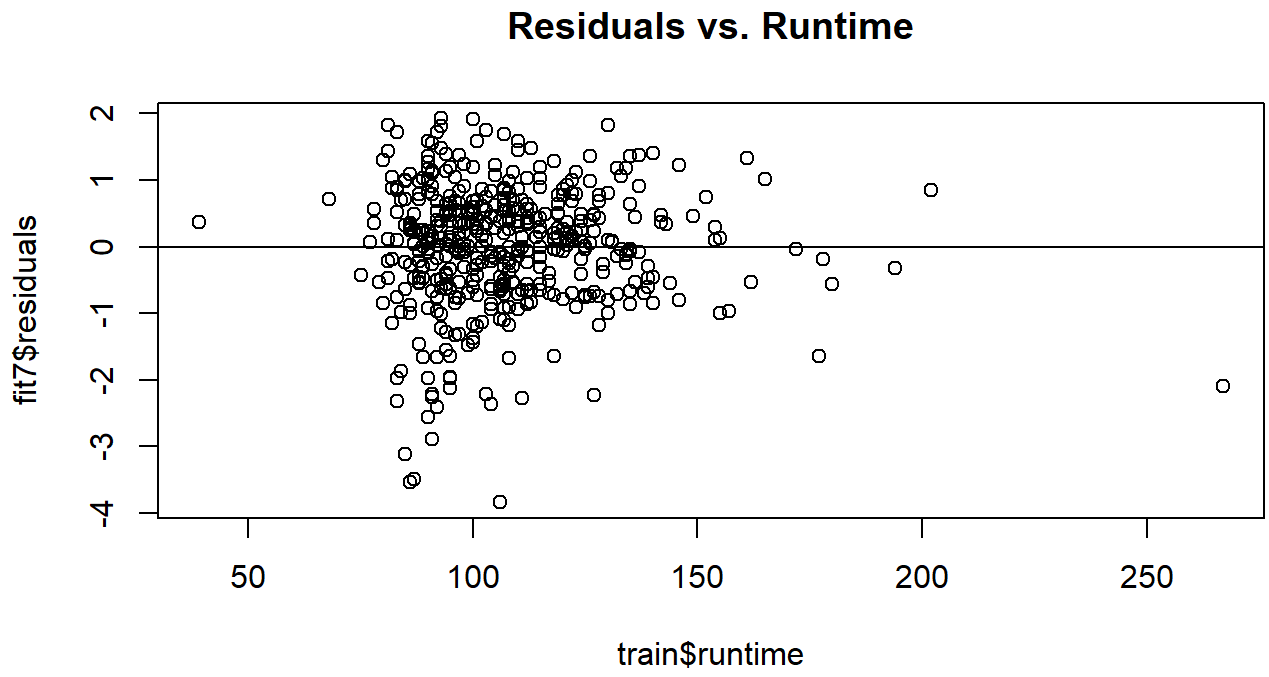
```
## lm(formula = imdb_rating ~ runtime + title_type + genre +
mpaa_rating +
##     best_dir_win, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8417 -0.5361  0.0922  0.5712  1.9267
##
## Coefficients:
##                                 Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                     6.777604   0.661082  10.252  <
2e-16 ***
## runtime                         0.012947   0.002283   5.671
2.6e-08 ***
## title_typeFeature Film         -1.101736   0.516249  -2.134
0.033392 *
## title_typeTV Movie             -2.059838   0.866887  -2.376
0.017927 *
## genreAnimation                 -0.866325   0.503790  -1.720
0.086214 .
## genreArt House & International  0.346577   0.331610   1.045
0.296541
## genreComedy                    -0.131135   0.188292  -0.696
0.486522
## genreDocumentary                0.548523   0.563593   0.973
0.330965
## genreDrama                      0.595487   0.158487   3.757
0.000195 ***
## genreHorror                    -0.078645   0.267731  -0.294
0.769091
## genreMusical & Performing Arts  0.290310   0.444798   0.653
0.514309
## genreMystery & Suspense         0.293135   0.208198   1.408
0.159857
## genreOther                      0.630494   0.330995   1.905
0.057460 .
## genreScience Fiction & Fantasy -0.583719   0.381082  -1.532
0.126313
## mpaa_ratingNC-17               -0.866065   0.744697  -1.163
0.245477
## mpaa_ratingPG                  -1.102021   0.352524  -3.126
0.001890 **
## mpaa_ratingPG-13               -1.283499   0.355889  -3.606
0.000346 ***
## mpaa_ratingR                   -0.957392   0.349164  -2.742
0.006359 **
## mpaa_ratingUnrated             -0.895016   0.406330  -2.203
```

```
0.028141 *
## best_dir_winyes                  0.526468   0.174517   3.017
0.002705 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9319 on 435 degrees of freedom
## Multiple R-squared:  0.3197, Adjusted R-squared:   0.29
## F-statistic: 10.76 on 19 and 435 DF,  p-value: < 2.2e-16
```

Model diagnosis

Check linear relationship

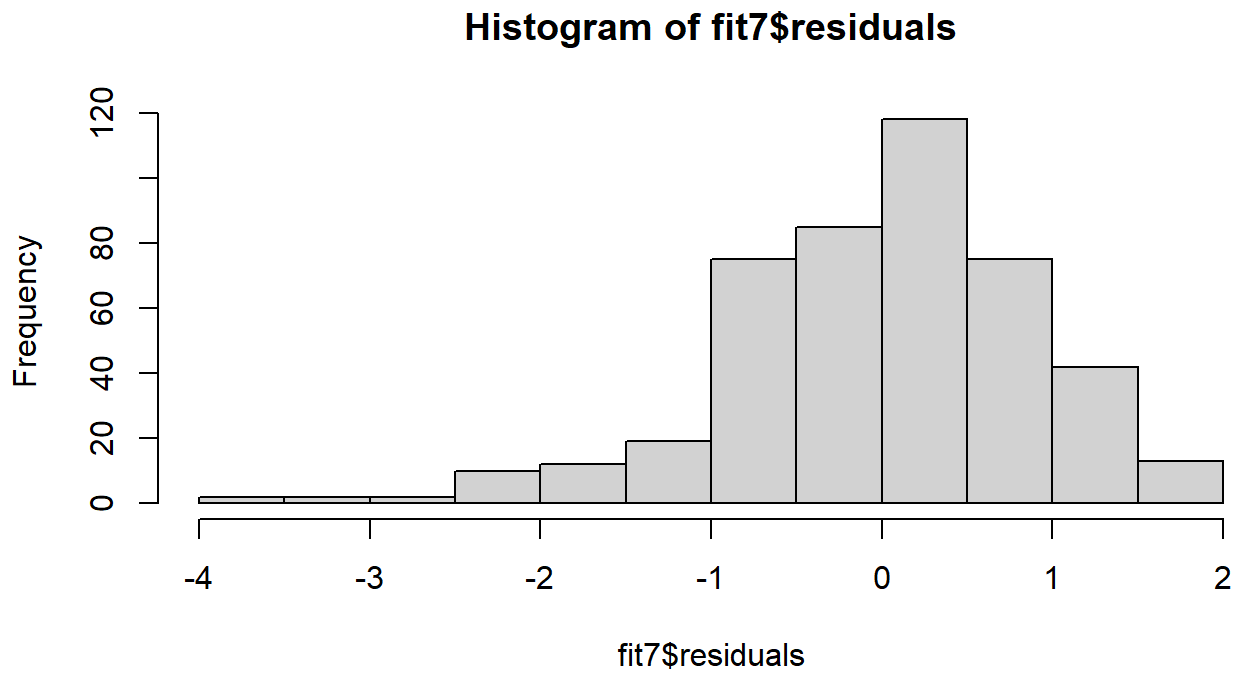
```
plot(fit7$residuals ~ train$runtime, main= "Residuals vs. Runtime")+
abline(h=0)
```



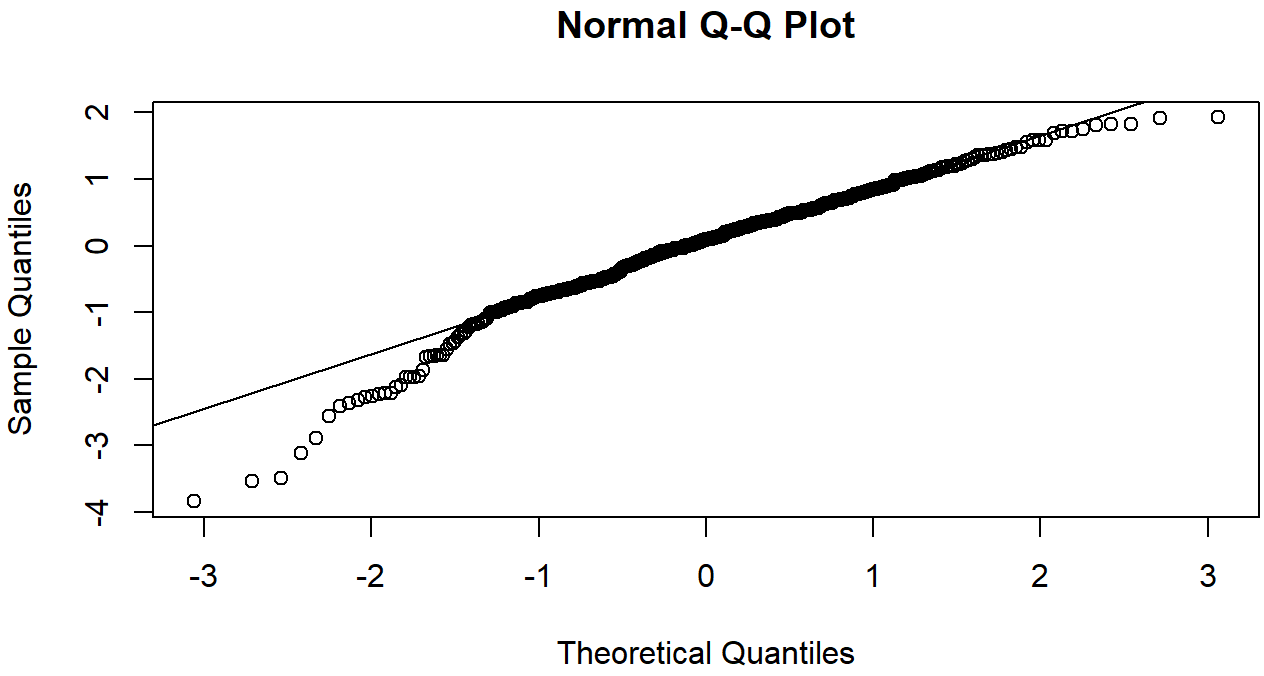

```
## integer(0)
```

Check nearly normal residual with mean 0

```
hist(fit7$residuals)
```

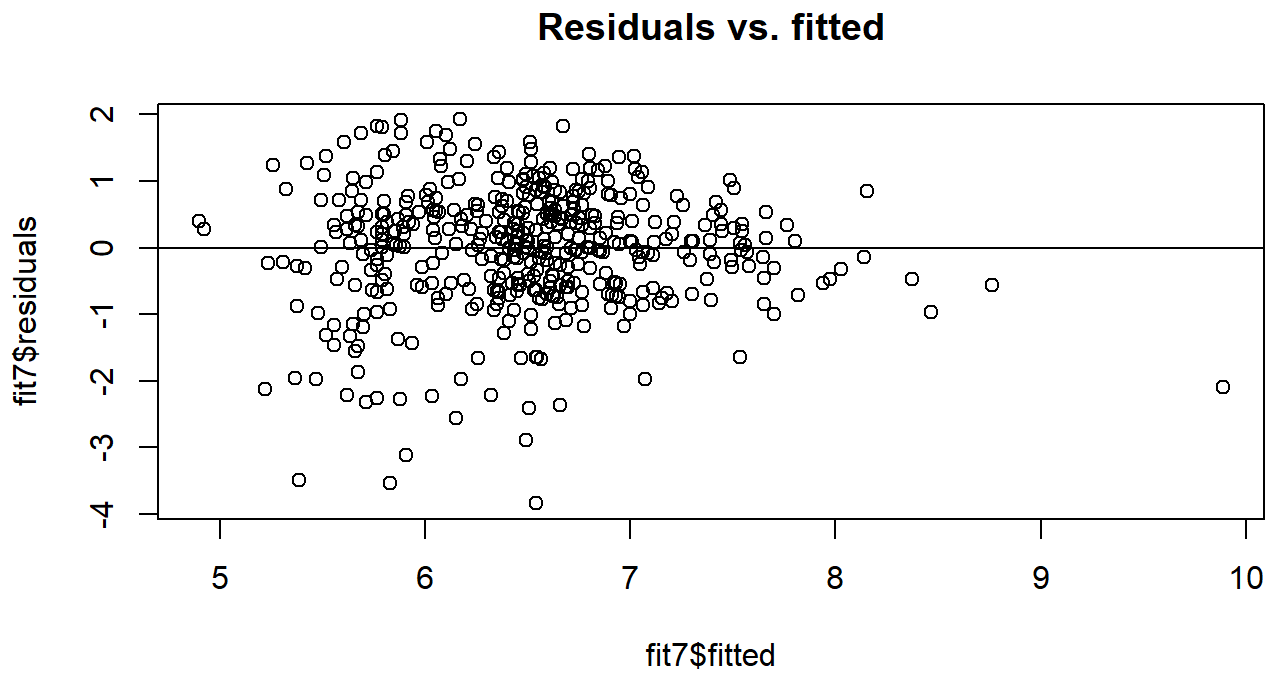



```
qqnorm(fit7$residuals)
qqline(fit7$residuals)
```




Check constant variability of residuals

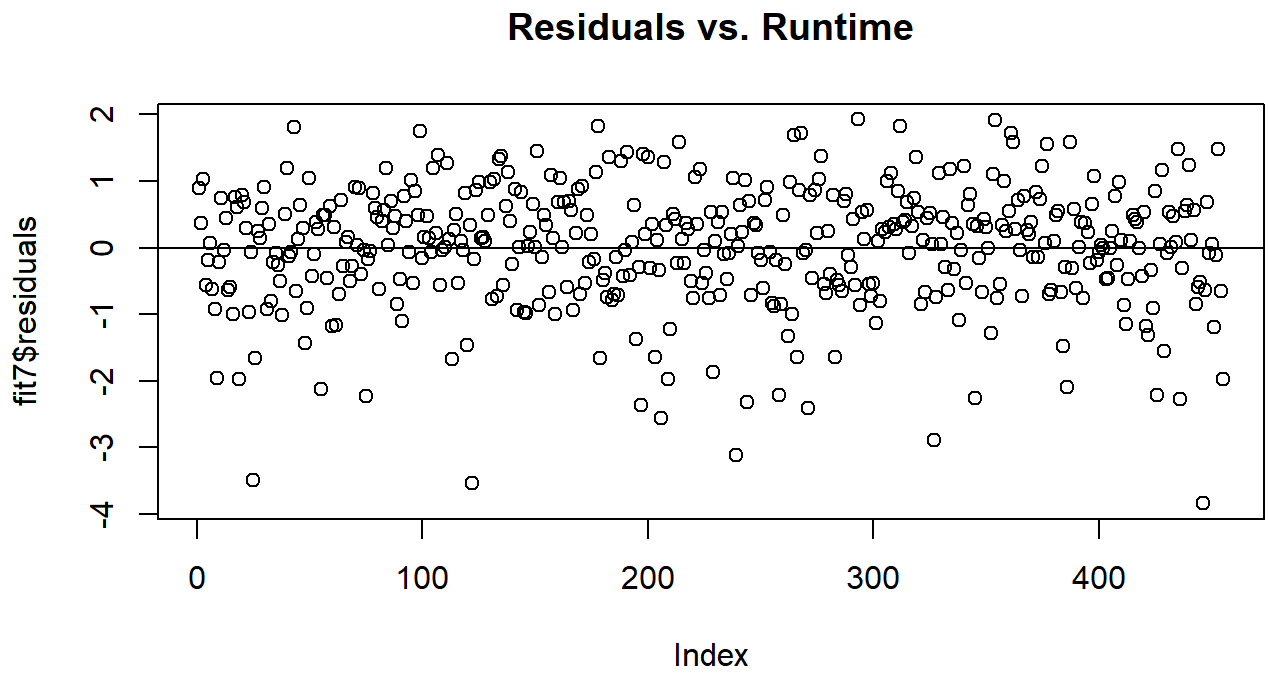
```
plot(fit7$residuals ~ fit7$fitted, main="Residuals vs. fitted")+
abline(h=0)
```




```
## integer(0)
```

Check independent residuals

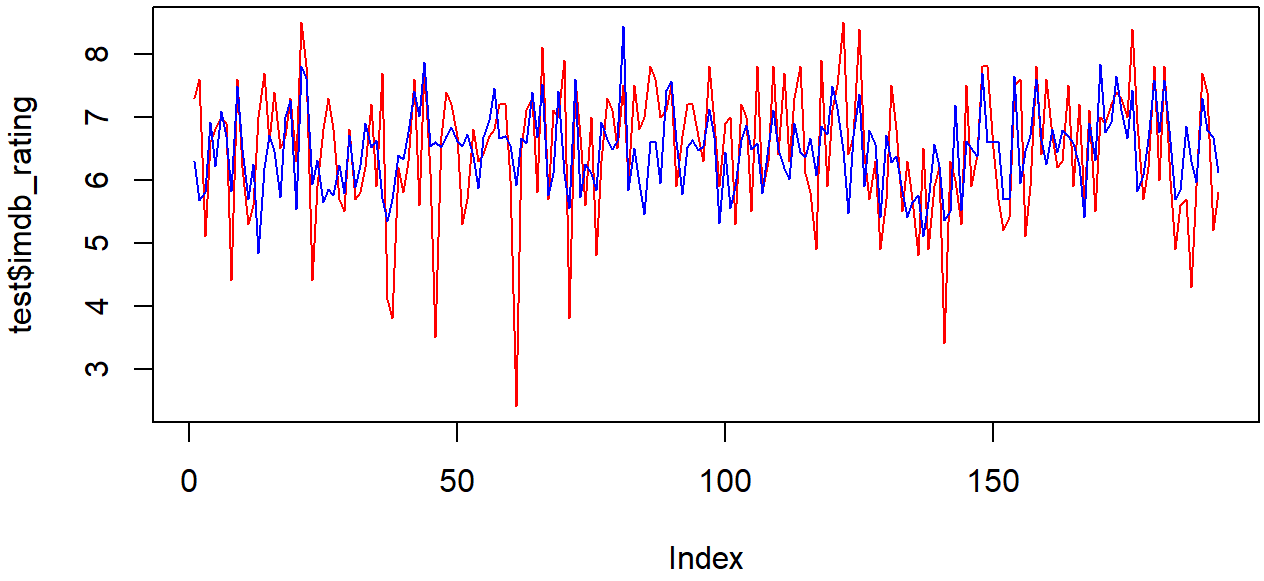
```
plot(fit7$residuals, main="Residuals vs. Runtime")+
abline(h=0)
```




```
## integer(0)
```

Comparing predicted vs actual value with test dataset

```
# getting predicted value
pred <- predict(fit7, test)

# predicted vs actual comparing visualisation
plot(test$imdb_rating,type = "l",lty=1.8,col="red")+
lines(pred,type = "l",lty=1.8,col="blue")
```




```
## integer(0)
```

```
#check accuracy with confidence interval
a<- predict(fit7, test, interval="prediction", level=0.95)
d <- as.data.frame(a)
b<- as.data.frame(d$lwr)
c<- as.data.frame(d$upr)

d_categorised <- ifelse(test$imdb_rating < c & test$imdb_rating > b,
"T","F")
table(d_categorised)
## d_categorised
##   F   T
##   8 184
```

```
184/192*100
## [1] 95.83333
```

The graph shows a moderate overlap between actual and predicted values. Also, about 96% of predicted imdb rating is within prediction interval.

---

Part 5: Prediction for a movie in 2016

Chosen movie: Sing (2016 American film) imdb rating: 7.1/10, runtime: 108 minutes, Feature film, comedy, mpaa rating: PG, best_dir_win: no,
(Source: https://www.imdb.com/title/tt3470600/)

```
dataadd <- data.frame(runtime=108, title_type="Feature Film",
genre="Comedy", mpaa_rating="PG", best_dir_win="no")

a<- predict(fit7, dataadd, interval="prediction", level=0.95)
```

Interpretation: with 95% confidence, imdb rating of Sing movie is expected to be between 4.1 and 7.7.

---

Part 6: Conclusion

As with given data for the project, imdb rating can be predicted by 5 variables that are runtime, title_type, genre, mpaa_rating, best_dir_win. The model has adjusted R2 of 0.29, which means 29% of the variation of imdb rating amongst movies are explained by independent variables. Though model found make correct prediction interval, the range of prediction interval, in other words, standard error of the prediction can be minimized by finding more meaningful variable and building a new model with higher adjusted R2. A model to predict revenue of a movie is also an valuable aspect that should be developed as it can tell how popular a movie is.