Modeling and prediction for movies

Setup

Load packages

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

Load data

Make sure your data and R Markdown files are in the same directory. When loaded your data file will be called `movies`. Delete this note when before you submit your work.

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

Part 1: Data

The data set we use for this analysis is **movies** data set which is comprised of 651 randomly sampled movies produced and released before 2016. Followings are the related variables.

variable	description
`audience_score`	Audience score on Rotten Tomatoes
`genre`	Genre of movie (Action & Adventure, Comedy, Documentary, Drama, Horror, Mystery & Suspense, Other)
`runtime`	Runtime of movie (in minutes)
`imdb_num_votes`	Number of votes on IMDB
`critics_score`	Critics score on Rotten Tomatoes
`top200_box`	Whether or not the movie is in the Top 200 Box Office list on BoxOfficeMojo (no, yes)
`thtr_rel_month`	Month the movie is released in theaters
`best_pic_nom`	Whether or not the movie was nominated for a best picture Oscar (no, yes)
`best_pic_win`	Whether or not the movie won a best picture Oscar (no, yes)

- **Generabizability**: According to the codebook, the data set is comprised of randomly sampled movies. The generalizability of this study is limited by the characteristics of the study movies. However, the analytics results of this study can be generalized to other movies with a large sample size and diverse genres or types.
- **Casualty**: As the data was gathered by observational study method rather than experiment, no causality relationship can be established.

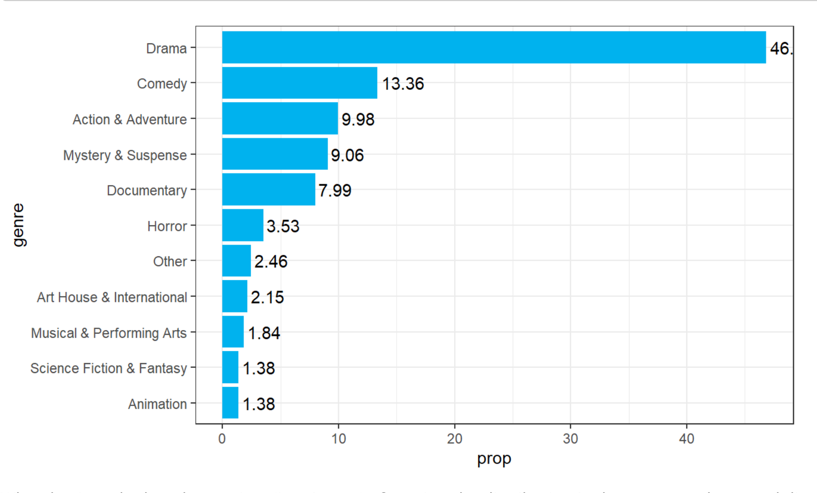
Part 2: Research question

The topic for this research analysis is about what attributes make a movie popular and find out the interesting things about movies. To specify, I am interested in whether varriables including genre, runtime, imdb_num_votes, critics_score and top200_box are significant predictors of audience score on Rotten Tomatoes.

Part 3: Exploratory data analysis

To begin with the genres of the observational movies, we can obviously see that *Drama* movies account for more than 46% of the movies collected, followed by Comedy (13.36%). Look at the genres, we recognize the popularity of Drama and Comedy movies while the animation genre movies have the lowest counts.

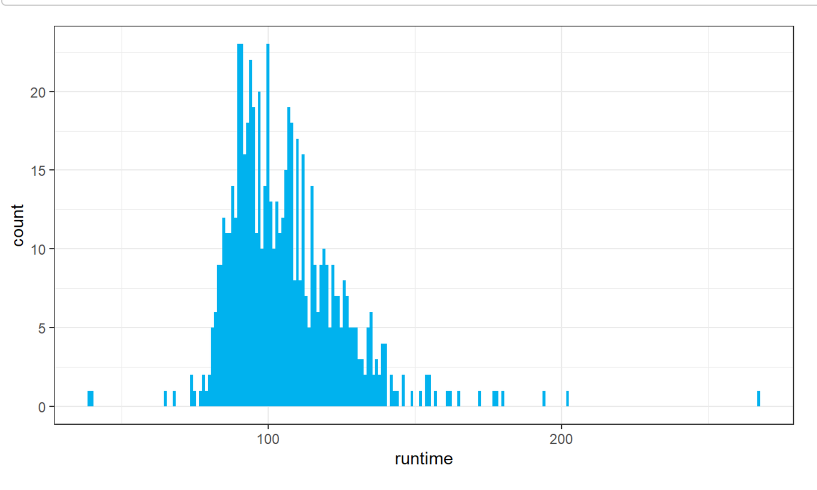
```
genre_sum <- movies %>% group_by(genre) %>% summarize(counts = n())
genre_sum <- genre_sum %>% mutate(prop = round(counts/sum(counts)*100, digits = 2))%>% arrange(prop)
# This trick update the factor levels
genre_sum <- genre_sum %>% mutate(genre=factor(genre, levels=genre))
ggplot(data=genre_sum, aes(x=genre, y=prop))+
  geom_bar(fill="deepskyblue2", stat = "identity") + coord_flip() +
  geom_text(aes(label = prop), hjust = -0.1)+
  theme_bw()
```



When having a look at the runtime (in minutes) of movies, the distribution is shown as nearly normal distribution with likely right-skewed form. The general runtime of the movies distributes around 100 minutes.

```
movies %>% ggplot(aes(x=runtime))+
  geom_histogram(fill="deepskyblue2", binwidth = 1)+
  theme_bw()
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



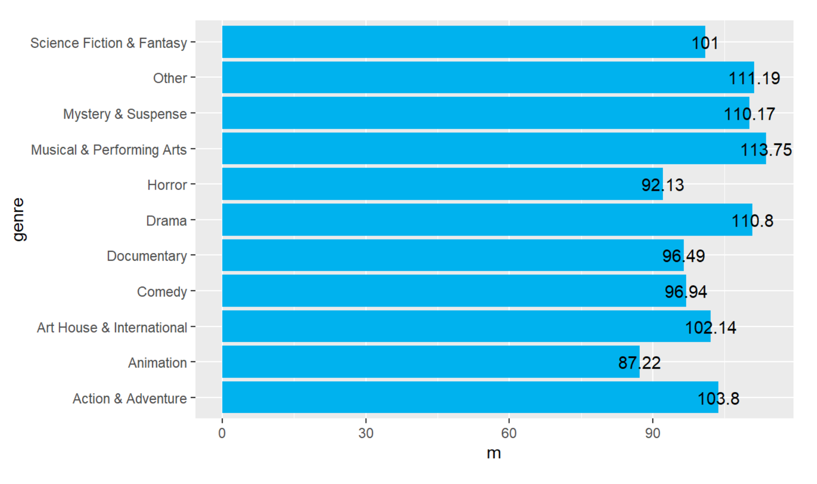
How are the distributions for different genres? Let's find out with some simple codes.

```
movies %>% ggplot(aes(x=runtime, group= genre, fill = genre))+
                geom_histogram(binwidth = 1)+
                facet_wrap(~genre)
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```

The means of genres' runtime do not show much difference. Animation movies have the lowest runtime mean at 87.22 minutes while musical and performing arts related movies length last most in minutes. There are seven over ten genres which have the duration lasting more than 100 minutes.

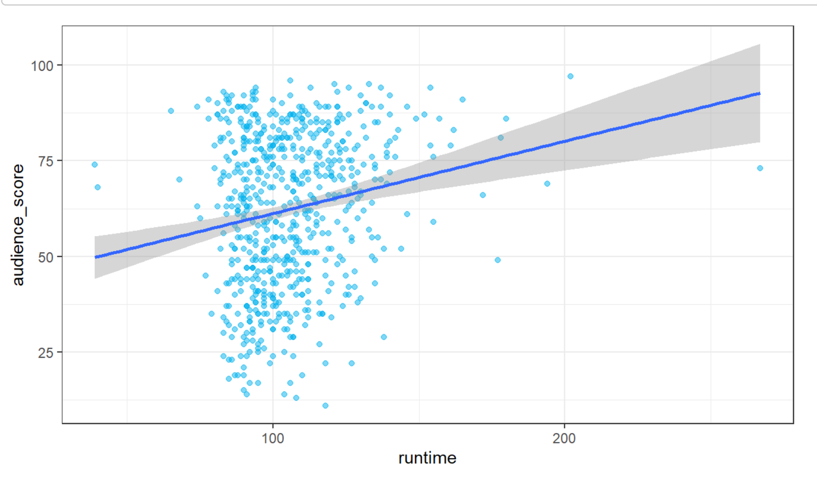
```
movies %>%
  group_by(genre) %>%
  summarize(m = mean(runtime, na.rm=T)) %>%
  ggplot(aes(x=genre, y=m, group= genre))+
  geom_bar(fill="deepskyblue2", stat="identity")+ coord_flip()+
  geom_text(aes(label = round(m, digits=2), hjust = +0.5))
```



Let us move on to association between the runtime of the movies and audience score on Rotten Tomatoes.

```
movies %>% ggplot(aes(x=runtime, y=audience_score))+
  geom_point(alpha=0.5,na.rm=T, color="deepskyblue2")+
  geom_smooth(method ='lm',na.rm=T)+
  theme_bw()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



In general, longer movies are positively associated with higher score from audiences.

We can have a look at the correlation between the numerical variables in interest.

```
ex_vars <- movies %>% select(runtime,imdb_num_votes,
                      critics_score)
ggpairs(ex_vars, na.rm=T)
```

```
## Warning in warn_if_args_exist(list(...)): Extra arguments: "na.rm" are being
## ignored. If these are meant to be aesthetics, submit them using the 'mapping'
## variable within ggpairs with ggplot2::aes or ggplot2::aes_string.
```

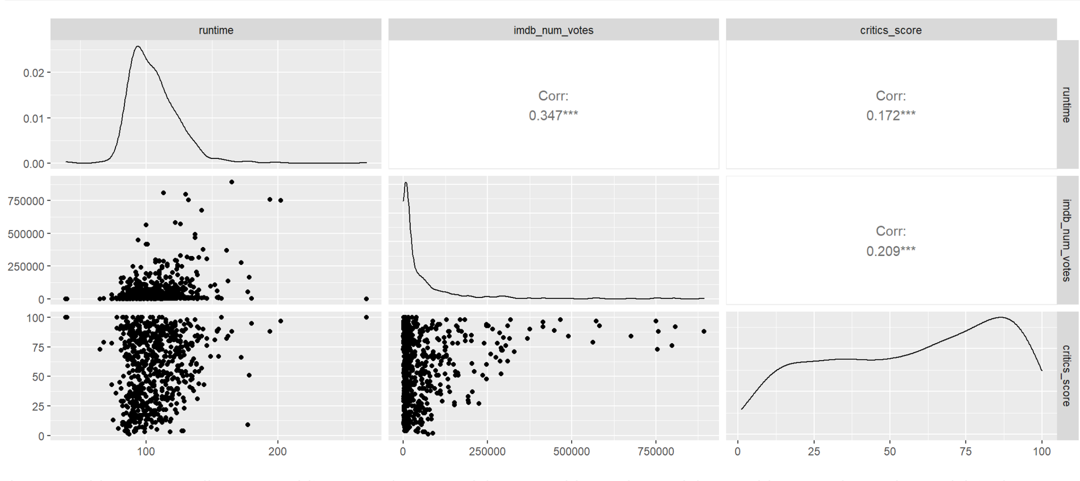
```
## Warning: Removed 1 rows containing non-finite values (stat_density).
```

```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removing 1 row that contained a missing value

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removing 1 row that contained a missing value
```

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

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



These variables are not collinear, so adding more than one of these variables to the model may add some value to the model. In this application and with these lowly-correlated predictors, it is reasonable to make use of these variables for linear regression model to predict audience scores.

Part 4: Modeling

To dig deep into the relationship between variables and figure out the effective model for good audience score prediction, we select variables and build an initial model with them.

Here we try 5 interested varriables including genre, runtime, imdb_num_votes, critics_score and top200_box for the initial model.

```
md <- lm(audience_score~genre+runtime+imdb_num_votes+
                      critics_score+top200_box+thtr_rel_month+
                      best_pic_win+best_pic_nom, data = movies)
summary(md)
```

```
##
## Call:
## lm(formula = audience_score ~ genre + runtime + imdb_num_votes +
##     critics_score + top200_box + thtr_rel_month + best_pic_win +
##     best_pic_nom, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.768  -9.089   0.325   9.167  43.715
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    3.337e+01  3.661e+00   9.113  < 2e-16 ***
## genreAnimation                 6.030e+00  4.856e+00   1.242  0.21484
## genreArt House & International  8.376e+00  4.022e+00   2.083  0.03767 *
## genreComedy                    2.146e-01  2.246e+00   0.096  0.92388
## genreDocumentary               1.311e+01  2.792e+00   4.696 3.25e-06 ***
## genreDrama                     3.215e+00  1.940e+00   1.658  0.09789 .
## genreHorror                   -7.135e+00  3.317e+00  -2.151  0.03184 *
## genreMusical & Performing Arts  1.360e+01  4.367e+00   3.115  0.00192 **
## genreMystery & Suspense       -3.732e+00  2.473e+00  -1.509  0.13171
## genreOther                     7.562e-01  3.834e+00   0.197  0.84371
## genreScience Fiction & Fantasy -6.603e+00  4.819e+00  -1.370  0.17108
## runtime                        1.330e-02  3.192e-02   0.417  0.67695
## imdb_num_votes                 3.453e-05  5.676e-06   6.083 2.04e-09 ***
## critics_score                  4.128e-01  2.190e-02  18.848  < 2e-16 ***
## top200_boxyes                 -9.911e-01  3.744e+00  -0.265  0.79133
## thtr_rel_month                -1.097e-01  1.557e-01  -0.705  0.48121
## best_pic_winyes               -8.316e+00  5.996e+00  -1.387  0.16591
## best_pic_nomyes                6.540e+00  3.522e+00   1.857  0.06375 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.53 on 632 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.5645, Adjusted R-squared:  0.5527
## F-statistic: 48.18 on 17 and 632 DF,  p-value: < 2.2e-16
```

As we can see in the summary statistics of the multiple linear regression, there are 2 significant predictors of audience score. They are the number of votes on IMDB and the critics score on Rotten Tomatoes with p-value < 0.05.

Followed the *backward elimination* approach using p-value criteria, we first remove top200_boxyes and refit the model. Again and again, we remove runtime,thtr_rel_month and best_pic_win to obtain the final model, since these variables do not bring the p-value <0.05 (They are not significant predictors of the audience score that we are finding).

```
md <- lm(audience_score~genre+imdb_num_votes+
                      critics_score+
                      best_pic_nom, data = movies)
summary(md)
```

```
##
## Call:
## lm(formula = audience_score ~ genre + imdb_num_votes + critics_score +
##     best_pic_nom, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.984  -8.714   0.106   9.275  43.653
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    3.404e+01  1.904e+00  17.882  < 2e-16 ***
## genreAnimation                 5.780e+00  4.811e+00   1.201  0.23001
## genreArt House & International  8.340e+00  4.009e+00   2.080  0.03788 *
## genreComedy                    4.016e-02  2.223e+00   0.018  0.98559
## genreDocumentary               1.287e+01  2.757e+00   4.669 3.70e-06 ***
## genreDrama                     3.297e+00  1.913e+00   1.724  0.08526 .
## genreHorror                   -7.262e+00  3.291e+00  -2.207  0.02767 *
## genreMusical & Performing Arts  1.363e+01  4.334e+00   3.145  0.00174 **
## genreMystery & Suspense       -3.609e+00  2.447e+00  -1.475  0.14076
## genreOther                     1.266e+00  3.812e+00   0.332  0.73991
## genreScience Fiction & Fantasy -6.640e+00  4.808e+00  -1.381  0.16778
## imdb_num_votes                 3.308e-05  5.243e-06   6.310 5.22e-10 ***
## critics_score                  4.129e-01  2.178e-02  18.956  < 2e-16 ***
## best_pic_nomyes                4.427e+00  3.155e+00   1.403  0.16107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.51 on 637 degrees of freedom
## Multiple R-squared:  0.5627, Adjusted R-squared:  0.5537
## F-statistic: 63.04 on 13 and 637 DF,  p-value: < 2.2e-16
```

Here we also find out the percentage of the variability of the audience score are explained by this model comprised of 4 above explanatory variables. For this model, 55% of the variability in audience score is explained by these variables.

With the coefficients table, we can write down the least squares regression line for the linear model:

$$\widehat{audience\_score} = (3.404e+01) + \ldots + (3.308e-05) \times imdb\_num\_votes + (4.129e-01) \times critics\_score + (4.427e+00) \times best\_pic\_nomyes$$

… is the brief for genre.

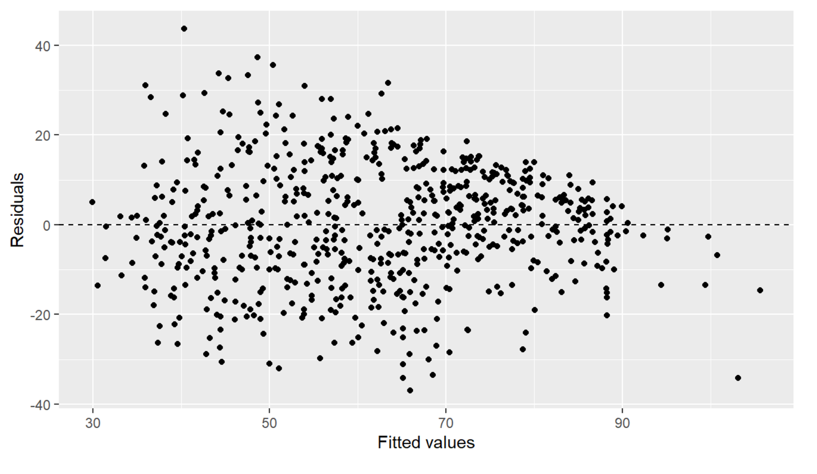
In the context of the relationship between audience score and these predictors, we can see that for each additional imdb_num_votes, the model predicts 3.308e-05 more audience score, on average.

Model diagnostics

To assess whether the linear model is reliable, we need to check for (1) linearity, (2) nearly normal residuals, and (3) constant variability.

Linearity: We should verify this condition with a plot of the residuals vs. fitted (predicted) values.

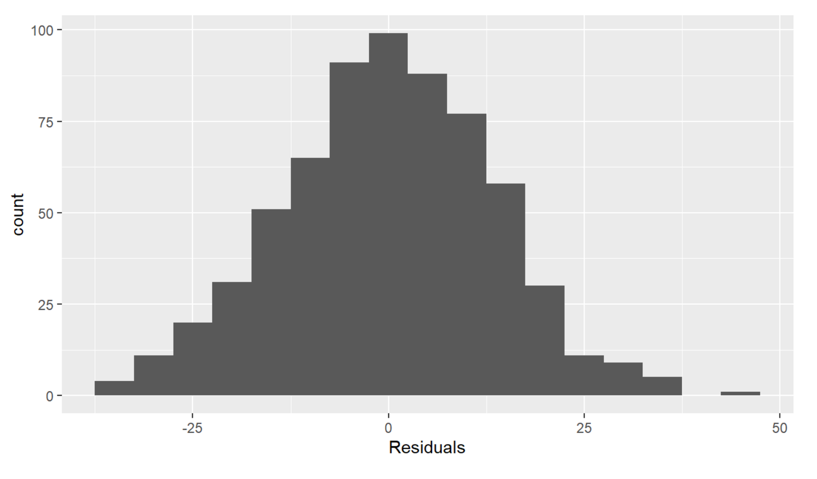
```
ggplot(data = md, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```



Obviously, the residuals appear to be randomly distributed around 0. The plot is also indicative of a linear relationship.

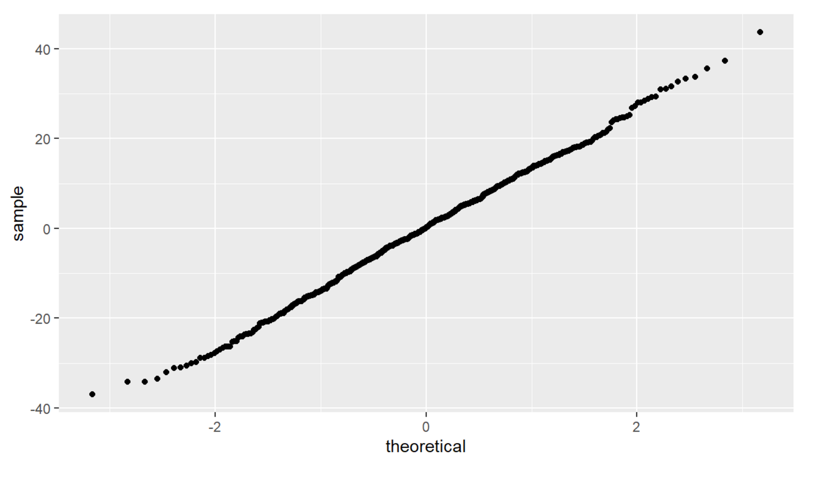
Nearly normal residuals: To check the condition, we can look at a histogram

```
ggplot(data = md, aes(x = .resid)) +
  geom_histogram(binwidth = 5) +
  xlab("Residuals")
```



or a normal probability plot of the residuals.

```
ggplot(data = md, aes(sample = .resid)) +
  stat_qq()
```



The residuals are fairly symmetric, centered at 0, with some outliers at the two tails, so it would be appropriate to deem the the normal distribution of residuals condition met.

Nearly normal residuals: Based on the previous residuals plot, we can see that the variablity of the residuals around the 0 line is roughly constant.

Part 5: Prediction

we want to use the model we created earlier to predict the evaluation score for the movie *Your Name* with the imdb_num_votes of 235,777, critics_score of 98, genre of Animation, not being nominated for a best picture Oscar.

Now we need to create a new data frame for this movie.

```
yourname <- data.frame(imdb_num_votes = 235777, genre = "Animation",
                       critics_score = 98, best_pic_nom="no" )
```

Then, I can do the prediction using the predict function:

```
predict(md, yourname)
```

```
##        1
## 88.09317
```

Actually, the audience score for Your Name is 94, our model's result is 88, it is quite cool.

We can also construct a prediction interval around this prediction, which will provide a measure of uncertainty around the prediction.

```
predict(md, yourname, interval = "prediction", level = 0.95)
```

```
##        fit      lwr      upr
## 1 88.09317 60.02571 116.1606
```

Hence, the model predicts, with 95% confidence, that the movie Your Name is expected to have an evaluation score between 60.02571 and 116.1606.

```
model9 <- lm(audience_score ~  genre + runtime + imdb_rating + critics_rating  + best_pic_nom, data = movies)
summary(model9)
```

```
##
## Call:
## lm(formula = audience_score ~ genre + runtime + imdb_rating +
##     critics_rating + best_pic_nom, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.051  -6.034   0.306   5.485  49.000
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -27.10755    4.11843  -6.582 9.74e-11 ***
## genreAnimation                  7.98262    3.49278   2.285 0.022615 *
## genreArt House & International  -0.39296    2.88477  -0.136 0.891692
## genreComedy                     1.72050    1.60736   1.070 0.284853
## genreDocumentary                1.22702    1.98923   0.617 0.537566
## genreDrama                     -0.02327    1.37212  -0.017 0.986477
## genreHorror                    -5.42190    2.37867  -2.279 0.022975 *
## genreMusical & Performing Arts   5.09212    3.11454   1.635 0.102556
## genreMystery & Suspense        -5.97481    1.77528  -3.366 0.000810 ***
## genreOther                      1.46545    2.75307   0.532 0.594708
## genreScience Fiction & Fantasy  -0.81921    3.48327  -0.235 0.814143
## runtime                        -0.04892    0.02216  -2.208 0.027614 *
## imdb_rating                    15.03688    0.50634  29.697  < 2e-16 ***
## critics_ratingFresh            -1.98041    1.12623  -1.758 0.079153 .
## critics_ratingRotten           -4.79082    1.27267  -3.764 0.000182 ***
## best_pic_nomyes                 3.35004    2.28073   1.469 0.142371
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.741 on 634 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7736, Adjusted R-squared:  0.7682
## F-statistic: 144.4 on 15 and 634 DF,  p-value: < 2.2e-16
```

Part 6: Conclusion

In conclusion, there are various contributors to make a movie popular and get the high audience score on Rotten Tomatoes. We have to mention genre, imdb_num_votes, critics_score and best_pic_nom as the good predictors for the effective linear model.

In this analysis, we still face up to some shortcomings like the limited data collection using random sampling method, which leads to the imbalance in the data we study. For example, Drama accounts for the largest proportion but we do not confirm that the imbalance is natural or due to the sampling technique. In reality, there are definitely many other factors that should be considered into account when examing the effection to the audience score. The variables in the given data set may be limited with some numerical and not be so diverse. To some extent, the sample dataset is good enough to develop a linear model to predict the interesting audience score with the acceptable accuracy.

@ This analysis is conducted by KhuongDT (data includes information from Rotten Tomatoes and IMDB)