Modeling and prediction for movies

Setup

Load packages

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

Load data

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

Part 1: Data

The dataset includes information from Rotten Tomatoes and IMDb for a random sample of movies produced and released before 2016.

Because random sampling was used, the result of the analysis can be generalized to the movies.

Because random assignment is not used, the result of the analysis cannot be used for causal inference.

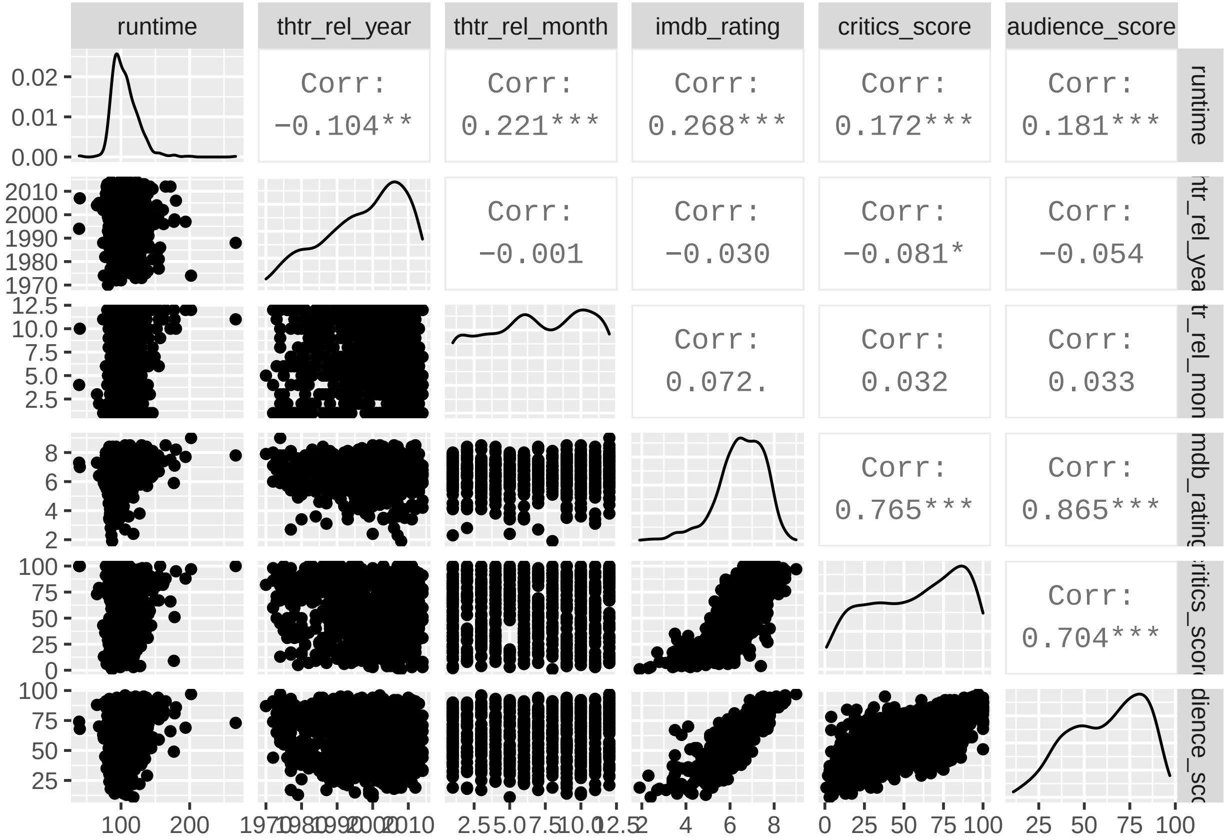
Part 2: Research question

What are the components of a popular movie?

Knowing the components of a popular movie can help movie producers to focus on the important things when making a movie, to improve the likelihood of the movie being popular.

Part 3: Exploratory data analysis

```r
ggpairs(movies, columns = c(4, 7, 8, 13, 16, 18))
```



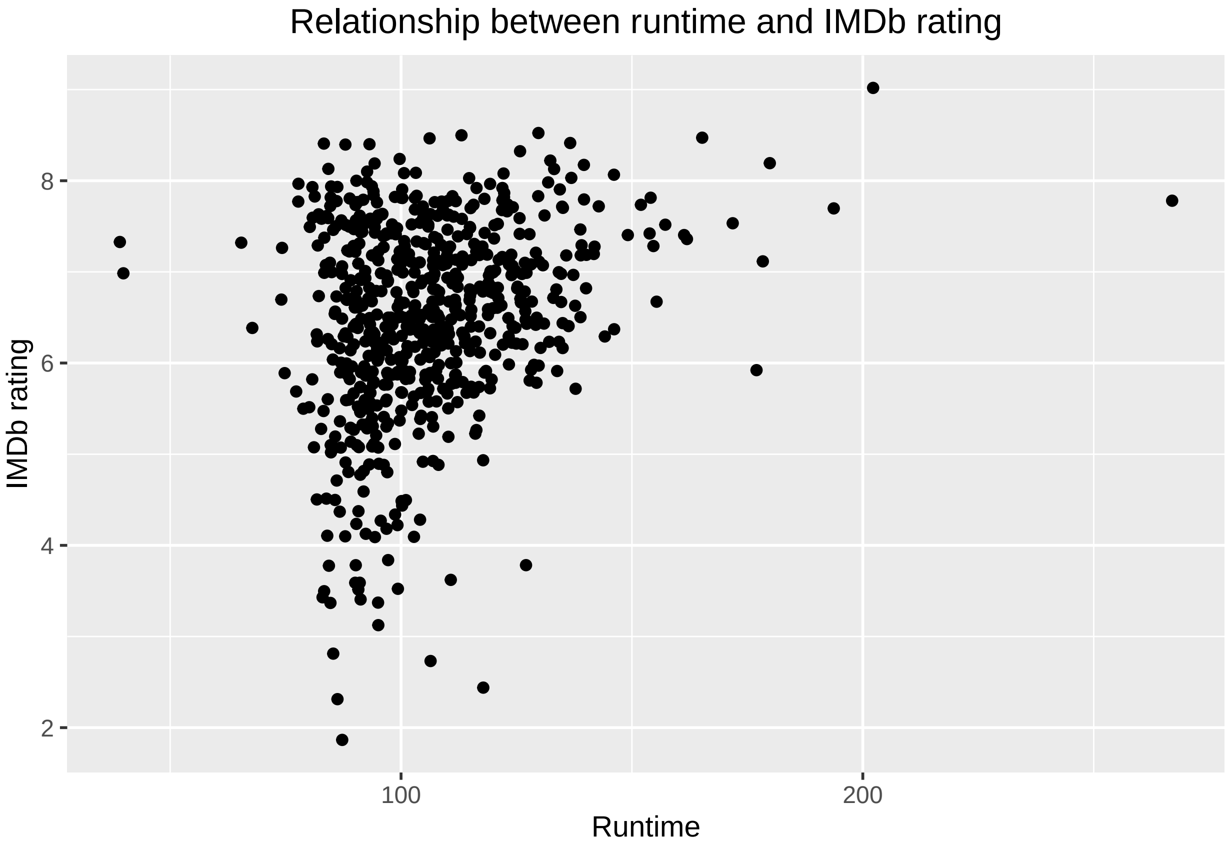
The pairs plot shows that audience_score and imdb_rating have strong linear relationship. I can use either of them as response variable of the model.

```
movies %>%
  group_by(genre) %>%
  summarise(lower_quartile = quantile(imdb_rating, probs = 0.25)
            , median_imdb_rating = median(imdb_rating)
            , upper_quartile = quantile(imdb_rating, probs = 0.75)
            ) %>%
  arrange(desc(median_imdb_rating))
```

```
## # A tibble: 11 x 4
##    genre                     lower_quartile median_imdb_rating upper_quartile
##    <fct>                              <dbl>              <dbl>          <dbl>
##  1 Documentary                         7.4                7.6            7.9
##  2 Musical & Performing Arts           6.95               7.55           7.72
##  3 Drama                               6.1                6.8            7.3
##  4 Other                               6.12               6.8            7.05
##  5 Art House & International           6.3                6.5            7.48
##  6 Mystery & Suspense                  6.05               6.5            7.05
##  7 Animation                           5.3                6.4            6.7
##  8 Action & Adventure                  5.6                6              6.7
##  9 Horror                              5.45               5.9            6.15
## 10 Science Fiction & Fantasy           5                  5.9            7.4
## 11 Comedy                              5.1                5.7            6.5
```

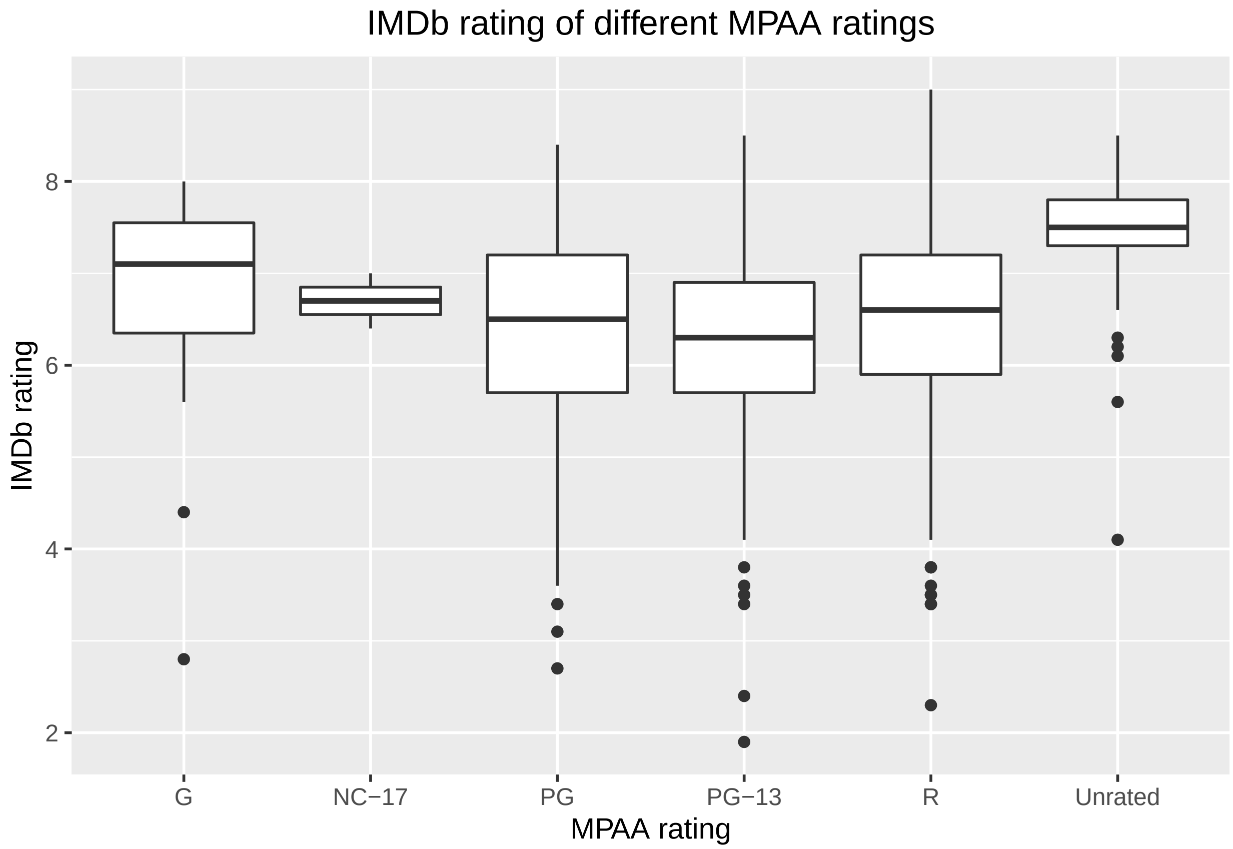
From this summary statistics, it shows that the genres “Musical & Performing Arts” and Documentary have much higher median IMDb rating than other genres.

```r
ggplot(data = movies, aes(x = runtime, y = imdb_rating)) +
  geom_jitter() +
  ggtitle('Relationship between runtime and IMDb rating') +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab('Runtime') +
  ylab('IMDb rating')
```



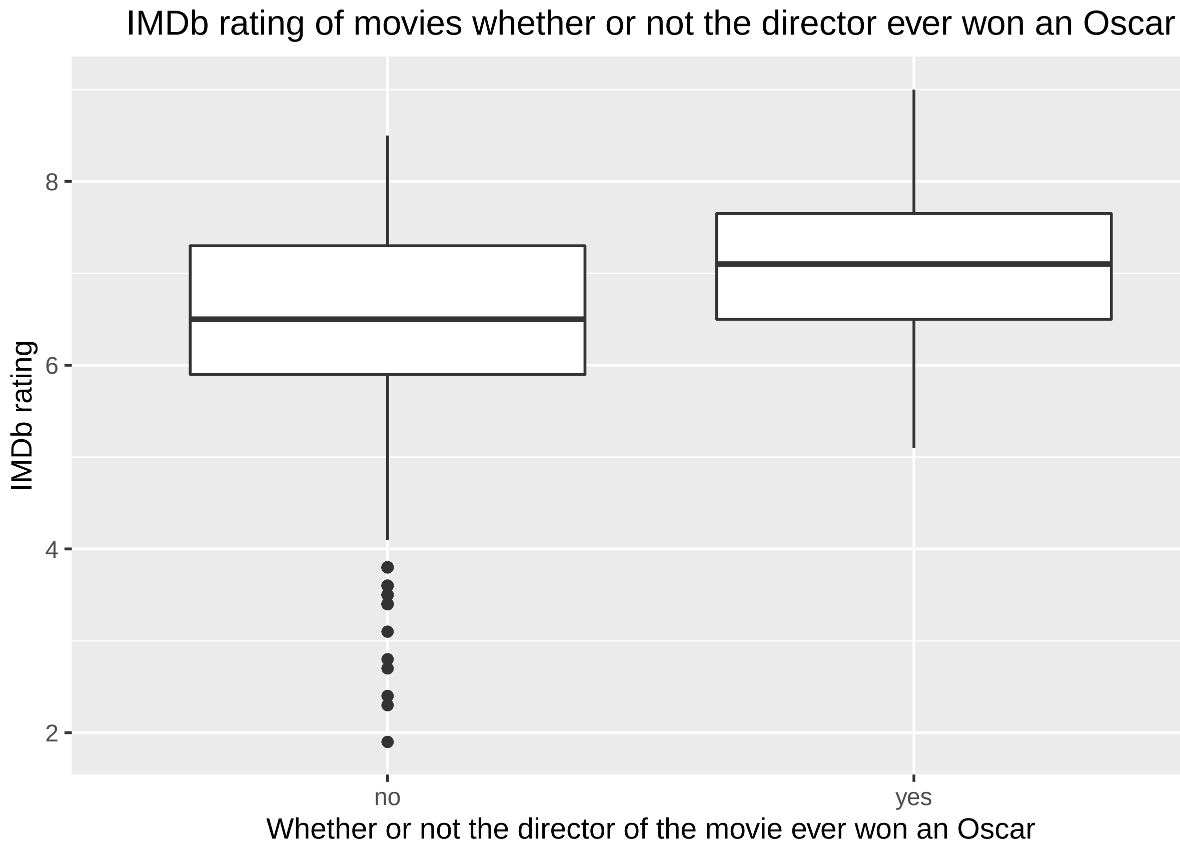
From this scatter plot, it shows that movies with runtime beyond 150 minutes is associated with higher IMDb rating.

```r
ggplot(data = movies, aes(x = mpaa_rating, y = imdb_rating)) +
  geom_boxplot() +
  ggtitle('IMDb rating of different MPAA ratings') +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab('MPAA rating') +
  ylab('IMDb rating')
```



From this box plot, it shows that the MPAA ratings “G” and “Unrated” have higher median IMDb rating than the other MPAA ratings.

```
ggplot(data = movies, aes(x = best_dir_win, y = imdb_rating)) +
  geom_boxplot() +
  ggtitle('IMDb rating of movies whether or not the director ever won an Oscar') +
  theme(plot.title = element_text(hjust = 0.5)) +
  ylab('IMDb rating') +
  xlab('Whether or not the director of the movie ever won an Oscar')
```



From this box plot, it shows that movies which the director ever won an Oscar having higher minimum IMDb rating and higher median IMDb rating.

Part 4: Modeling

Variables To Consider For Full Model I will consider the following variables for the full model:

- response variable
  - imdb_rating
- explanatory variables
  - genre
  - runtime
  - mpaa_rating
  - thtr_rel_month
  - best_dir_win

Model Selection Because I am interested in knowing the components of a popular movie, I will use the p-value approach for model selection to find out which variables are significant. I will use backwards elimination as it will take shorter time than forward selection when doing the p-value approach.

```
# re-level so that the coefficient of each level will be easier to interpret
movies$mpaa_rating <- relevel(movies$mpaa_rating, ref = 'PG-13')
movies$genre <- relevel(movies$genre, ref = 'Animation')
```

Step 1

```r
m_movies <- lm(imdb_rating ~ genre + runtime + mpaa_rating +
                 thtr_rel_month + best_dir_win
               , data = movies)
summary(m_movies)
```

```
## 
## Call:
## lm(formula = imdb_rating ~ genre + runtime + mpaa_rating + thtr_rel_month + 
##     best_dir_win, data = movies)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.8924 -0.5122  0.0543  0.6055  1.9723 
## 
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                    4.030321   0.412171   9.778  < 2e-16 ***
## genreAction & Adventure        0.281477   0.357335   0.788 0.431161    
## genreArt House & International 0.904942   0.427433   2.117 0.034637 *  
## genreComedy                    0.214707   0.359737   0.597 0.550824    
## genreDocumentary               1.933433   0.373709   5.174 3.09e-07 ***
## genreDrama                     0.900273   0.352758   2.552 0.010942 *  
## genreHorror                    0.172610   0.398425   0.433 0.664995    
## genreMusical & Performing Arts 1.432288   0.437740   3.272 0.001126 ** 
## genreMystery & Suspense        0.690406   0.369245   1.870 0.061977 .  
## genreOther                     0.852808   0.416531   2.047 0.041032 *  
## genreScience Fiction & Fantasy 0.016806   0.453403   0.037 0.970444    
## runtime                        0.012871   0.002112   6.094 1.92e-09 ***
## mpaa_ratingG                   0.926748   0.256722   3.610 0.000331 ***
## mpaa_ratingNC-17               0.452235   0.654254   0.691 0.489680    
## mpaa_ratingPG                  0.255360   0.118912   2.147 0.032136 *  
## mpaa_ratingR                   0.307062   0.097496   3.149 0.001713 ** 
## mpaa_ratingUnrated             0.458931   0.186336   2.463 0.014047 *  
## thtr_rel_month                 0.006009   0.010486   0.573 0.566799    
## best_dir_winyes                0.406282   0.150036   2.708 0.006955 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.9156 on 631 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3074, Adjusted R-squared:  0.2877 
## F-statistic: 15.56 on 18 and 631 DF,  p-value: < 2.2e-16
```

Because the variable thtr_rel_month is not significant (i.e. p-value above significance level 0.05), and has the highest p-value, I will remove it from the model and refit the smaller model.

Step 2

```r
m_movies <- lm(imdb_rating ~ genre + runtime + mpaa_rating + best_dir_win
               , data = movies)
summary(m_movies)
```

```
## 
## Call:
## lm(formula = imdb_rating ~ genre + runtime + mpaa_rating + best_dir_win, 
##     data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8921 -0.5035  0.0686  0.6068  1.9788
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      4.045954   0.411049   9.843  < 2e-16 ***
## genreAction & Adventure          0.273593   0.356880   0.767  0.44359
## genreArt House & International   0.897364   0.427001   2.102  0.03599 *
## genreComedy                      0.210617   0.359475   0.586  0.55815
## genreDocumentary                 1.926737   0.373328   5.161 3.29e-07 ***
## genreDrama                       0.891670   0.352251   2.531  0.01160 *
## genreHorror                      0.165741   0.398033   0.416  0.67726
## genreMusical & Performing Arts   1.426836   0.437404   3.262  0.00117 **
## genreMystery & Suspense          0.677776   0.368390   1.840  0.06626 .
## genreOther                       0.838196   0.415529   2.017  0.04410 *
## genreScience Fiction & Fantasy   0.007460   0.452869   0.016  0.98686
## runtime                          0.013149   0.002055   6.399 3.04e-10 ***
## mpaa_ratingG                     0.930417   0.256506   3.627  0.00031 ***
## mpaa_ratingNC-17                 0.447490   0.653854   0.684  0.49398
## mpaa_ratingPG                    0.260684   0.118485   2.200  0.02816 *
## mpaa_ratingR                     0.311612   0.097121   3.208  0.00140 **
## mpaa_ratingUnrated               0.459276   0.186236   2.466  0.01392 *
## best_dir_winyes                  0.408043   0.149925   2.722  0.00667 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9151 on 632 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3071, Adjusted R-squared:  0.2884
## F-statistic: 16.47 on 17 and 632 DF,  p-value: < 2.2e-16
```

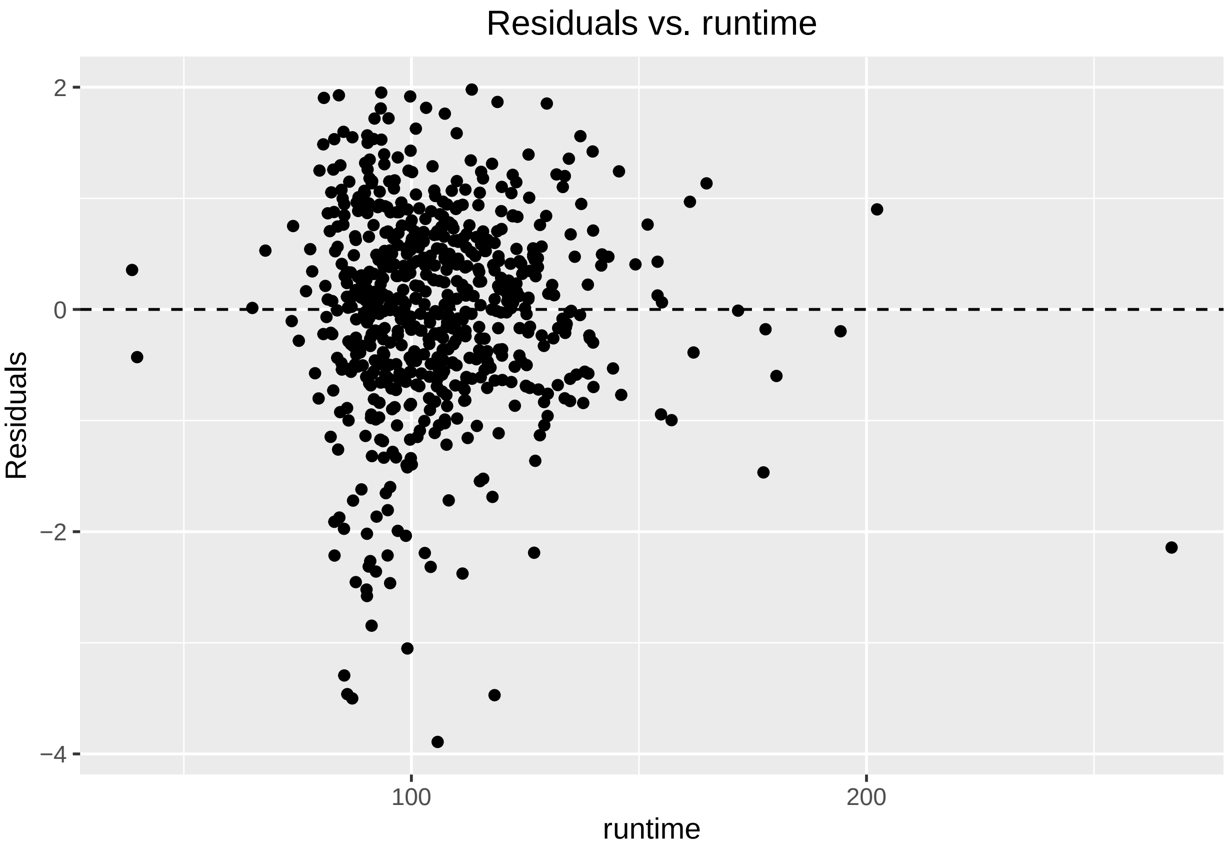
Because all the 4 variables are significant (i.e. p-value below significance level 0.05), I will stop here.

Model Diagnostics I will now check whether the following conditions are met for the multiple linear regression model to be valid:

- Linear relationships between (numerical) x and y
- Nearly normal residuals with mean 0
- Constant variability of residuals
- Independent residuals

Linear relationships between (numerical) x and y

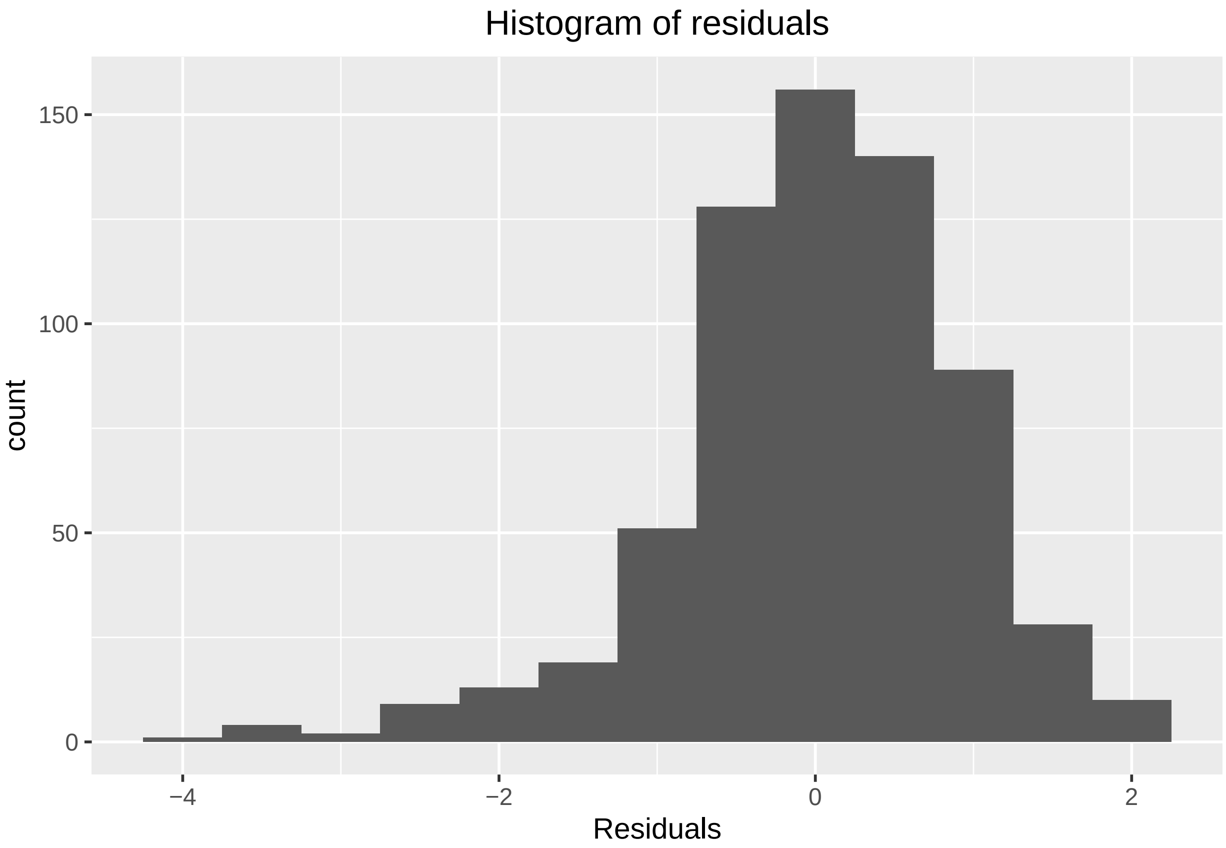
```r
filtered_movies <- movies %>%
  filter(!is.na(genre), !is.na(runtime))

ggplot(data = filtered_movies
       , aes(x = runtime, y = m_movies$residuals)) +
  geom_jitter() +
  geom_hline(yintercept = 0, linetype = 'dashed') +
  ggtitle('Residuals vs. runtime') +
  theme(plot.title = element_text(hjust = 0.5)) +
  ylab('Residuals')
```



From this residuals plot, it shows that there are random scatter around 0, which means that the numerical explanatory variable is linearly related to the response variable.

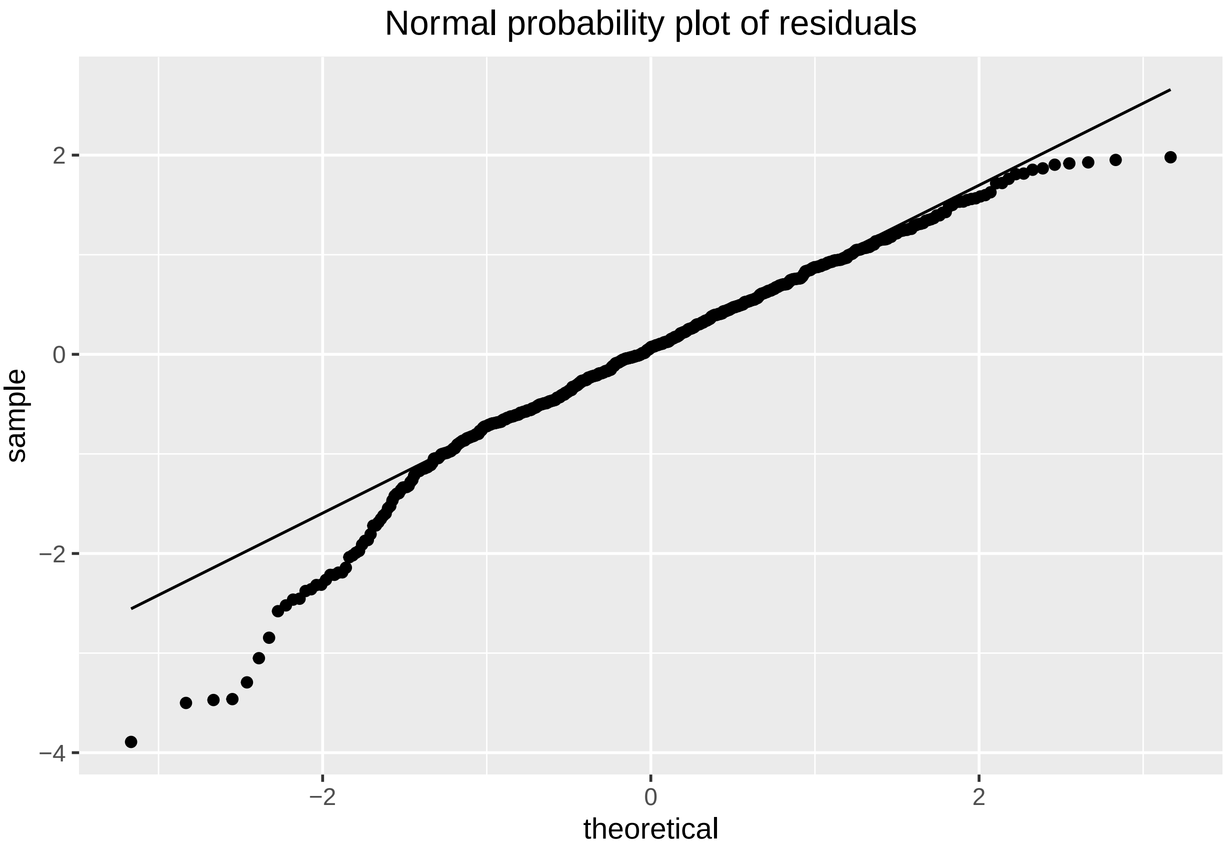
Nearly normal residuals with mean 0

```
ggplot(data = m_movies, aes(x = .resid)) +
  geom_histogram(binwidth = 0.5) +
  ggtitle('Histogram of residuals') +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab('Residuals')
```



From the histogram, it shows that the distribution has a slight skew to the left, but quite normal and centered at 0.

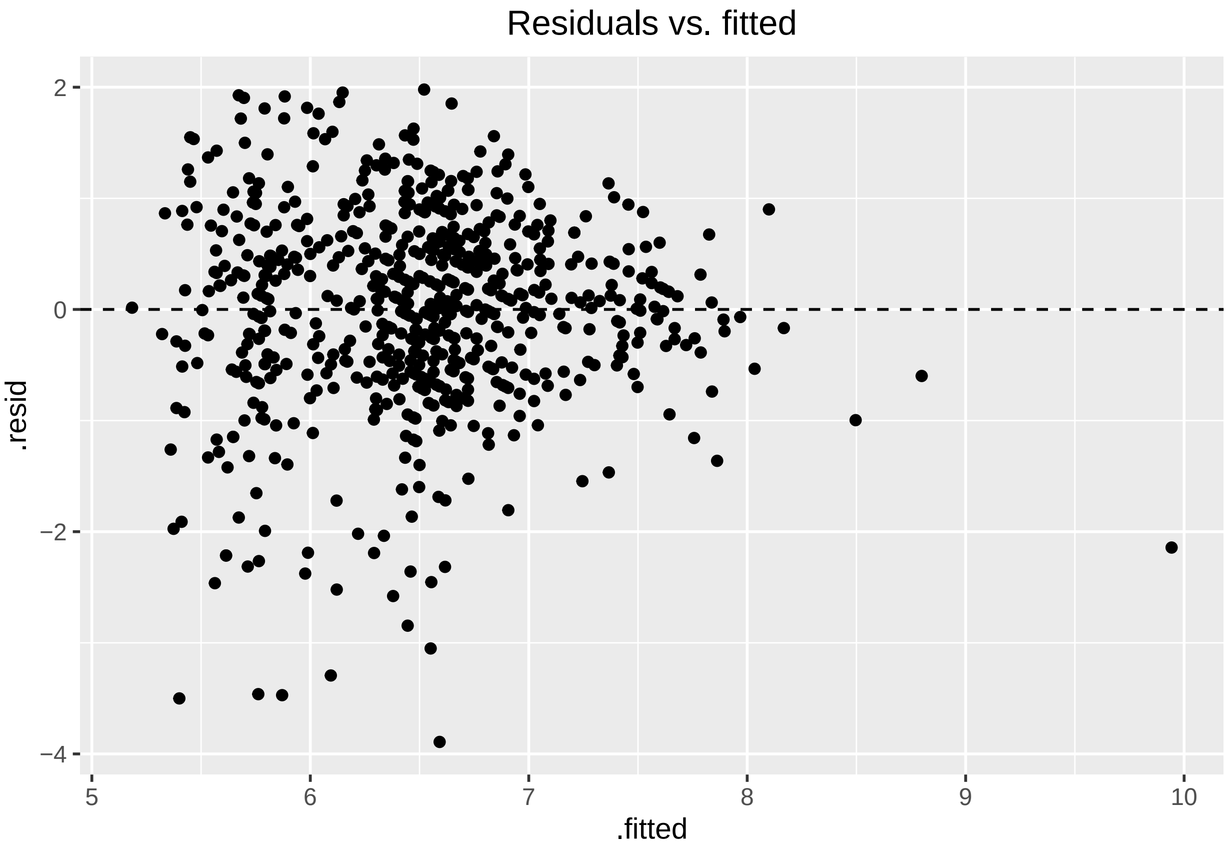
```r
ggplot(data = m_movies, aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line() +
  ggtitle('Normal probability plot of residuals') +
  theme(plot.title = element_text(hjust = 0.5))
```



From the normal probability plot, other than the slight deviation at the left tail, the line looks mostly straight, indicating that the residuals are mostly normally distributed.

Constant variability of residuals

```
ggplot(data = m_movies, aes(x = .fitted, y = .resid)) +
  geom_jitter() +
  geom_hline(yintercept = 0, linetype = 'dashed') +
  ggtitle('Residuals vs. fitted') +
  theme(plot.title = element_text(hjust = 0.5))
```



From the plot, it shows that the random scatter of residuals are concentrated around 0. Other than the few residuals below -2, the variability looks constant around 0.

Independent residuals

```
movies %>%
  group_by(thtr_rel_year) %>%
  summarise(count = n()) %>%
  arrange(desc(thtr_rel_year)) %>%
  top_n(n = 10)
```

```
## # A tibble: 10 x 2
##    thtr_rel_year count
##            <dbl> <int>
##  1          2012    27
##  2          2011    26
##  3          2008    22
##  4          2007    33
##  5          2006    33
##  6          2004    28
##  7          2003    25
##  8          2002    23
##  9          1996    27
## 10          1993    22
```

The data are random sampled, and the sizes of sample without replacement each year are less than 10% of the number of movies released each year, therefore the observations are independent.

```r
summary(m_movies)
```

Interpretation Of Model Coefficients

```
##
## Call:
## lm(formula = imdb_rating ~ genre + runtime + mpaa_rating + best_dir_win,
##     data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8921 -0.5035  0.0686  0.6068  1.9788
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    4.045954   0.411049   9.843  < 2e-16 ***
## genreAction & Adventure        0.273593   0.356880   0.767  0.44359
## genreArt House & International 0.897364   0.427001   2.102  0.03599 *
## genreComedy                    0.210617   0.359475   0.586  0.55815
## genreDocumentary               1.926737   0.373328   5.161 3.29e-07 ***
## genreDrama                     0.891670   0.352251   2.531  0.01160 *
## genreHorror                    0.165741   0.398033   0.416  0.67726
## genreMusical & Performing Arts 1.426836   0.437404   3.262  0.00117 **
## genreMystery & Suspense        0.677776   0.368390   1.840  0.06626 .
## genreOther                     0.838196   0.415529   2.017  0.04410 *
## genreScience Fiction & Fantasy 0.007460   0.452869   0.016  0.98686
## runtime                        0.013149   0.002055   6.399 3.04e-10 ***
## mpaa_ratingG                   0.930417   0.256506   3.627  0.00031 ***
## mpaa_ratingNC-17               0.447490   0.653854   0.684  0.49398
## mpaa_ratingPG                  0.260684   0.118485   2.200  0.02816 *
## mpaa_ratingR                   0.311612   0.097121   3.208  0.00140 **
## mpaa_ratingUnrated             0.459276   0.186236   2.466  0.01392 *
## best_dir_winyes                0.408043   0.149925   2.722  0.00667 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9151 on 632 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3071, Adjusted R-squared:  0.2884
## F-statistic: 16.47 on 17 and 632 DF,  p-value: < 2.2e-16
```

All else held constant, for each 1 minute increase in the runtime of movie, the model predicts the IMDb rating to be higher on average by 0.013.

All else held constant, the model predicts that movie with genre “Documentary” have IMDb rating 1.93 higher than movie with genre “Animation” (i.e. reference level), on average.

All else held constant, the model predicts that movie with genre “Art House & International” have IMDb rating 0.90 higher than movie with genre “Animation” (i.e. reference level), on average.

All else held constant, the model predicts that movie with genre “Drama” have IMDb rating 0.89 higher than movie with genre “Animation” (i.e. reference level), on average.

All else held constant, the model predicts that movie with MPAA rating of “G” have IMDb rating 0.93 higher than movie with MPAA rating of “PG-13” (i.e. reference level), on average.

All else held constant, the model predicts that movie which the director ever won an Oscar have IMDb rating 0.41 higher than movie which the director never won an Oscar, on average.

Part 5: Prediction

I will pick a movie that is not in the sample and do prediction of its IMDb rating using the model I developed.

I picked the movie "Captain America: Civil War", which can be found from the following URL:

- https://www.imdb.com/title/tt3498820/?ref_=adv_li_tt

```
movies$genre %>% unique()
```

```
##  [1] Drama                          Comedy
##  [3] Horror                         Documentary
##  [5] Action & Adventure             Art House & International
##  [7] Musical & Performing Arts Mystery & Suspense
##  [9] Animation                      Science Fiction & Fantasy
## [11] Other
## 11 Levels: Animation Action & Adventure Art House & International ... Science Fiction & Fantasy
```

Its genre from the website is "Action, Adventure, Sci-Fi", so I will use "Action & Adventure" as the genre for prediction.

```
movies %>%
  filter(director %in% c('Anthony Russo', 'Joe Russo')) %>%
  select(director, best_dir_win)
```

```
## # A tibble: 0 x 2
## # ... with 2 variables: director <chr>, best_dir_win <fct>
```

Neither of the 2 directors of this movie ever won as Oscar, so I will use "no" as input for best_dir_win.

```
new_movie <- data.frame(genre = 'Action & Adventure'
                        , runtime = 147
                        , mpaa_rating = 'PG-13'
                        , best_dir_win = 'no')

predict(m_movies, new_movie, interval = 'prediction'
        , level = 0.95)
```

```
##        fit      lwr      upr
## 1 6.252429 4.429415 8.075444
```

The model predicts, with 95% confidence, that the movie "Captain America: Civil War" with genre "Action & Adventure", with runtime 147 minutes, with MPAA rating "PG-13", and with directors who never won an Oscar is expected to have a IMDb rating between 4.43 and 8.08.

The actual IMDb rating of this movie from the website is 7.8, which is within the prediction interval.

Part 6: Conclusion

Regarding the components of a popular movie, what I have found is that audience prefer watching movies with long runtime, and movies that are suitable for audience of all ages (i.e. MPAA rating "G"). The top 3 genres that audience prefer are "Documentary", followed by "Musical & Performing Arts" and "Drama". Also, audience like watching movies that are produced by directors who ever won an Oscar.

The shortcoming of this study is that the model has R^2 of 0.3071, meaning that only 30.71% of the variability of the response variable (i.e. IMDb rating) can be explained by the model.