Movies' Rating Modeling and Prediction

Setup

Load packages

```
library(ggplot2)
library(dplyr)
library(statsr)
library(graphics)
```

Load data

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

Part 1: Data

As indicated in the project files, this data set is comprised of 651 randomly sampled movies produced and released before 2016. It is thus an observational study with random sampling. The results of this study can probably be generalized but no causality can be established as there was no random assignment used.

One reservation that one may do relates to the fact that films produced and released since 2016 are not included in the sample and this may affect the conclusions about the population.

Part 2: Research question

The research question is set in the project files and relates to the attributes that make a movie popular. In other words, we are expected to establish an association between the attributes of a movie and its score.

Even though no causality can be established, it is still important for the movie-making industry to know what factors are associated with its popularity.

Part 3: Exploratory data analysis

Step 1. Variables selection and data clean up.

Creating a data set with relevant variables: `title_type`, `genre`, `runtime`, `mpaa_rating`, `thtr_rel_month`, `thtr_rel_day`, `dvd_rel_month`, `dvd_rel_day`, `critics_score`, `imdb_num_votes`, `best_actor_win`, `best_actress_win`, `best_dir_win` (explanatory variables) and `imdb_rating`, `audience_score` (response variables)

```r
movies2 <- movies %>%

    # selecting variables
    select(title_type, genre, runtime, mpaa_rating, thtr_rel_month, thtr_rel_day,
           dvd_rel_month, dvd_rel_day, critics_score, imdb_num_votes,
           best_actor_win, best_actress_win, best_dir_win,
           imdb_rating, audience_score) %>%

    # excluding "Unrated" movies from `mpaa_rating`
    filter(mpaa_rating != "Unrated")
```

Converting release date variables from numerical to categorical

```r
# converting `thtr_rel_month`, `thtr_rel_day`, `dvd_rel_month`, `dvd_rel_day` to categorical
    movies2$thtr_rel_month <- as.factor(movies2$thtr_rel_month)
    movies2$thtr_rel_day <- as.factor(movies2$thtr_rel_day)
    movies2$dvd_rel_month <- as.factor(movies2$dvd_rel_month)
    movies2$dvd_rel_day <- as.factor(movies2$dvd_rel_day)
```

Using relevant variables `imdb_rating` and `audience_score` to create the response variable for the model (calculated as the average of two original scores).

```r
# mutating a new variable
movies2 <- movies2 %>%
    mutate(pop = (imdb_rating + audience_score)/2)
```

Excluded variables:

Variable	Comments
`studio`	levels are almost as numerous as observations
`thtr_rel_year`	cannot be used for prediction as it is a past event that will never repeat
`dvd_rel_year`	same
`critics_rating`	already reflected in `critics_score`
`audience_rating`	already reflected in `audience_score`
`best_pic_nom`	cannot be used since we are measuring popularity among the audience
`best_pic_win`	same
`top200_box`	cannot be used since can be affected by advertisement expenses and other confounding variables

Variable	Comments
`director`	choice of the director is reflected in `best_dir_win` variable
`actor1`	casting is reflected in `best_actor_win`, `best_actress_win` variables
`actor2`	same
`actor3`	same
`actor4`	same
`actor5`	same
`imdb_url`	variable provided for information purposes only
`rt_url`	same

A special remark should be made on `imdb_num_votes` since the number of votes a movie receives can be treated as both explanatory and response variable. In this research we treat it as an explanatory variable.

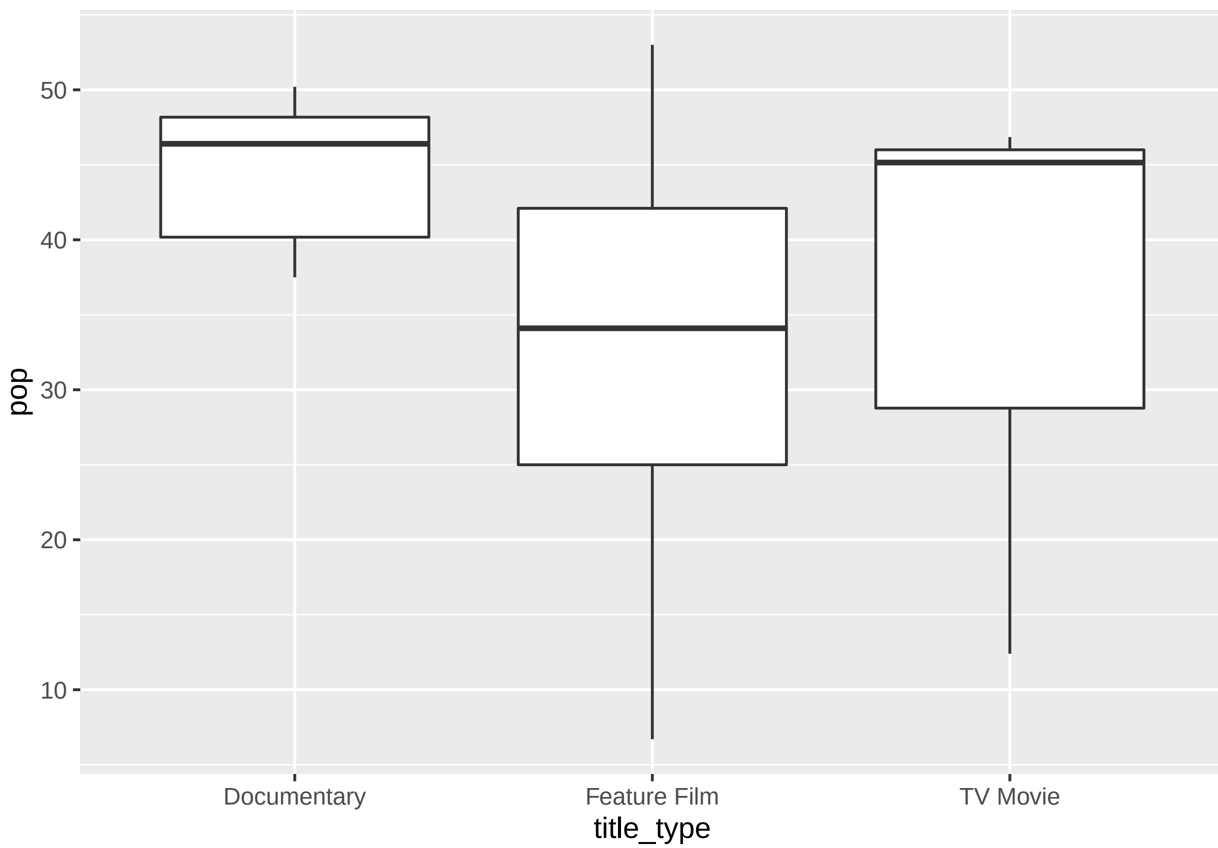
Step 2. Looking for collinearity between the explanatory variables.

We will account for collinearity when we build the MLR model using the Adjusted R-squared selection method as if a variable adds no new information to the model (is collinear) it will be dropped during the model selection stage.

Step 3. EDA and simple regression for some of the variables.

Case 1. Considering the relationship between `title_type` (categorical) and the response variable `pop` (numerical).

```
# creating a plot
movies2 %>%
    ggplot(aes(x = title_type, y = pop)) +
        geom_boxplot()
```



Mean of the pop variable for Feature films looks different from the mean score of Documentaries and TV movies.

Summary statistics

```
# mean popularity broken down by movie type
movies2 %>%
    group_by(title_type) %>%
    summarise(mean_dd = mean(pop)) %>%
    arrange(desc(mean_dd))
```

```
## # A tibble: 3 x 2
##   title_type   mean_dd
##   <fct>          <dbl>
## 1 Documentary     44.7
## 2 TV Movie        34.8
## 3 Feature Film    33.3
```

Summary statistics provide the same result as the plot.

Simple linear regression for categorical data

```r
# regression model for `title_type` and `pop`
slr1 <- lm(pop ~ title_type, data = movies2)
summary(slr1)
```

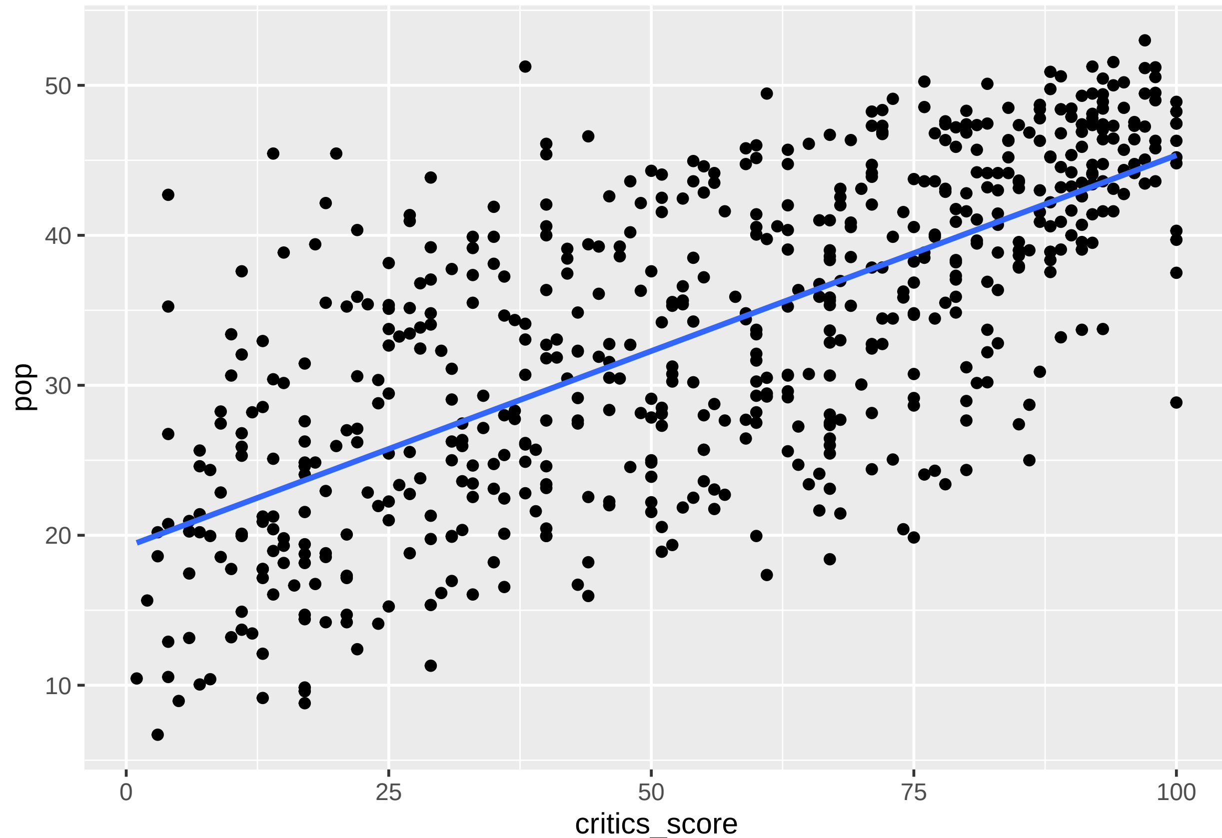
```
##
## Call:
## lm(formula = pop ~ title_type, data = movies2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.5746  -7.8246   0.9254   8.6254  19.7254
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)              44.722      2.142  20.875  < 2e-16 ***
## title_typeFeature Film  -11.447      2.185  -5.239 2.24e-07 ***
## title_typeTV Movie       -9.922      6.307  -1.573    0.116
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.27 on 598 degrees of freedom
## Multiple R-squared:  0.04394,    Adjusted R-squared:  0.04074
## F-statistic: 13.74 on 2 and 598 DF,  p-value: 1.461e-06
```

From the box plot, summary statistics and the simple regression model for categorical data we can conclude that while R-squared is small and there is no significant difference between a TV movie and a Documentary there is a significant difference between the reference level (Documentary) and a Feature film, the model itself has a very small p-value and appears to be statistically significant.

Case 2. Consider the relationship between `critics_score` (numerical) and the explanatory variable `pop` (numerical).

```r
# creating a plot and a trenline
movies2 %>%
    ggplot(aes(x = critics_score, y = pop)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE)
```

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



The plot shows a positive linear relationship.

Using a simple regression model

```
# regression model for `critics_score` and `pop`
slr2 <- lm(pop ~ critics_score, data = movies2)
summary(slr2)
```

```
##
## Call:
## lm(formula = pop ~ critics_score, data = movies2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.9613  -5.0170   0.1658   5.3224  22.5636
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    19.2315     0.6782   28.36   <2e-16 ***
## critics_score   0.2611     0.0109   23.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.505 on 599 degrees of freedom
## Multiple R-squared:  0.489,  Adjusted R-squared:  0.4882
```

```
## F-statistic: 573.3 on 1 and 599 DF,  p-value: < 2.2e-16
```

Correlation coefficient

```r
# correlation between 'critics_score' and 'pop'
cor(movies2$critics_score,movies2$pop)
```

```
## [1] 0.6993016
```

From the scatter plot, the trend line, the simple regression model and the correlation coefficient output we can conclude that there is a significant association between popularity and critics score. At the same time, R-squared is modest showing that only half of the response variable variation can be explained by the `critics_score` variable.

Conclusion

EDA on pairs of variables (one explanatory variable and the response variable) showed that there is a significant correlation between at least one pair of variables. At the same time, the coefficients of determination show that a significant amount of variation is not explained by the above factors. We will try to improve R-squared and proceed to an MLR model.

Part 4: Modeling

Variables selection. Variables selection for the full model and reasoning for excluding some of the variables *are given in Part 3.*

Model selection method. In this research we are going to use a *forward selection with adjusted R-squared* approach as it provides more reliable predictions than the p-value and does not depend on the choice of the significance level cutoff.

Forward selection with adjusted R-squared

Step	Variables included	Adjusted R-squared
Step 1	pop ~ title_type	0.04074
	pop ~ genre	0.133
	pop ~ runtime	0.05369
	pop ~ mpaa_rating	0.01012
	pop ~ thtr_rel_month	-0.006472
	pop ~ thtr_rel_day	-0.003599
	pop ~ dvd_rel_month	0.0009197
	pop ~ dvd_rel_day	0.003385
	pop ~ critics_score	**0.4882**
	pop ~ imdb_num_votes	0.1177
	pop ~ best_actor_win	0.0001963
	pop ~ best_actress_win	0.0008094

Step	Variables included	Adjusted R-squared
	pop ~ best_dir_win	0.01292
Step 2	pop ~ critics_score + title_type	0.4893
	pop ~ critics_score + genre	0.5068
	pop ~ critics_score + runtime	0.496
	pop ~ critics_score + mpaa_rating	0.4862
	pop ~ critics_score + thtr_rel_month	0.4831
	pop ~ critics_score + thtr_rel_day	0.4787
	pop ~ critics_score + dvd_rel_month	0.4894
	pop ~ critics_score + dvd_rel_day	0.4762
	pop ~ critics_score + imdb_num_votes	**0.5158**
	pop ~ critics_score + best_actor_win	0.4874
	pop ~ critics_score + best_actress_win	0.4874
	pop ~ critics_score + best_dir_win	0.4874
Step 3	pop ~ critics_score + imdb_num_votes + title_type	0.5209
	pop ~ critics_score + imdb_num_votes + genre	**0.5434**
	pop ~ critics_score + imdb_num_votes + runtime	0.5163
	pop ~ critics_score + imdb_num_votes + mpaa_rating	0.5149
	pop ~ critics_score + imdb_num_votes + thtr_rel_month	0.5111
	pop ~ critics_score + imdb_num_votes + thtr_rel_day	0.5078
	pop ~ critics_score + imdb_num_votes + dvd_rel_month	0.5153
	pop ~ critics_score + imdb_num_votes + dvd_rel_day	0.5052
	pop ~ critics_score + imdb_num_votes + best_actor_win	0.5153
	pop ~ critics_score + imdb_num_votes + best_actress_win	0.5156
	pop ~ critics_score + imdb_num_votes + best_dir_win	0.5154
Step 4	pop ~ critics_score + imdb_num_votes + genre + title_type	0.5421
	pop ~ critics_score + imdb_num_votes + genre + runtime	0.5433
	pop ~ critics_score + imdb_num_votes + genre + mpaa_rating	0.5431
	pop ~ critics_score + imdb_num_votes + genre + thtr_rel_month	0.5383
	pop ~ critics_score + imdb_num_votes + genre + thtr_rel_day	0.536
	pop ~ critics_score + imdb_num_votes + genre + dvd_rel_month	0.54
	pop ~ critics_score + imdb_num_votes + genre + dvd_rel_day	0.5315
	pop ~ critics_score + imdb_num_votes + genre + best_actor_win	0.5428
	pop ~ critics_score + imdb_num_votes + genre + best_actress_win	0.5433
	pop ~ critics_score + imdb_num_votes + genre + best_dir_win	0.5428

Final MLR model output

```r
# final MLR model
m_final <- lm(pop ~ critics_score + imdb_num_votes + genre, data = movies2)
summary(m_final)
```

```
##
## Call:
## lm(formula = pop ~ critics_score + imdb_num_votes + genre, data = movies2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.1808  -4.5230   0.1313   4.8854  22.8542
##
## Coefficients:
```

```
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1.919e+01  9.962e-01  19.261  < 2e-16 ***
## critics_score                    2.225e-01  1.166e-02  19.085  < 2e-16 ***
## imdb_num_votes                   1.846e-05  2.657e-06   6.949 9.78e-12 ***
## genreAnimation                   2.757e+00  2.525e+00   1.092  0.27523
## genreArt House & International   4.011e+00  2.415e+00   1.661  0.09734 .
## genreComedy                      8.031e-03  1.166e+00   0.007  0.99451
## genreDocumentary                 6.052e+00  1.868e+00   3.240  0.00126 **
## genreDrama                       1.864e+00  1.004e+00   1.857  0.06381 .
## genreHorror                     -3.384e+00  1.785e+00  -1.896  0.05843 .
## genreMusical & Performing Arts   7.961e+00  2.445e+00   3.256  0.00120 **
## genreMystery & Suspense         -1.695e+00  1.284e+00  -1.320  0.18735
## genreOther                       3.680e-01  2.047e+00   0.180  0.85742
## genreScience Fiction & Fantasy -3.577e+00  2.523e+00  -1.418  0.15675
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.088 on 588 degrees of freedom
## Multiple R-squared:  0.5526, Adjusted R-squared:  0.5434
## F-statistic: 60.51 on 12 and 588 DF,  p-value: < 2.2e-16
```

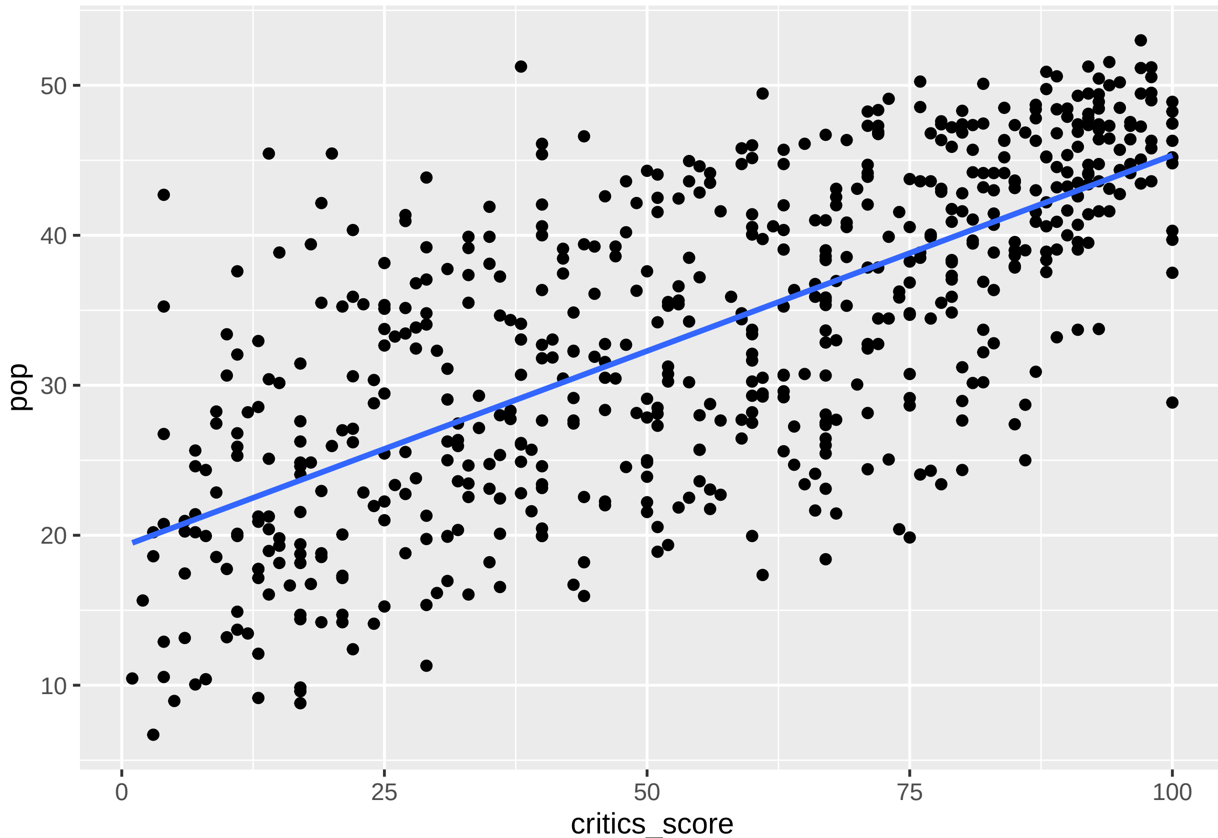
Model diagnostics

1. Linear relationship between the numerical x and y.

We have two numerical variables: `critics_score` and `imdb_num_votes`

```r
# creating a plot and a trenline
movies2 %>%
    ggplot(aes(x = critics_score, y = pop)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE)
```

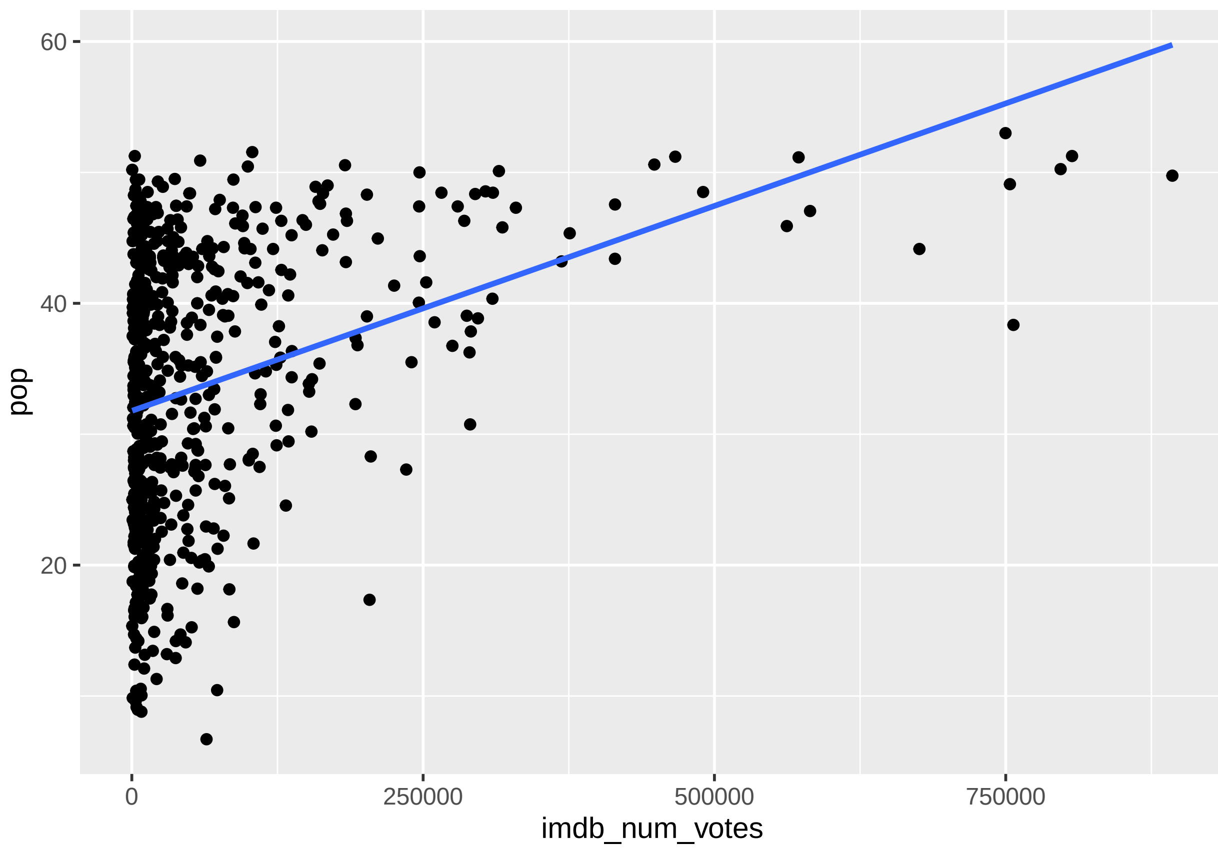
```
## `geom_smooth()` using formula 'y ~ x'
```



The data appear to have a linear relationship.

```
# creating a plot and a trenline
movies2 %>%
    ggplot(aes(x = imdb_num_votes, y = pop)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE)
```

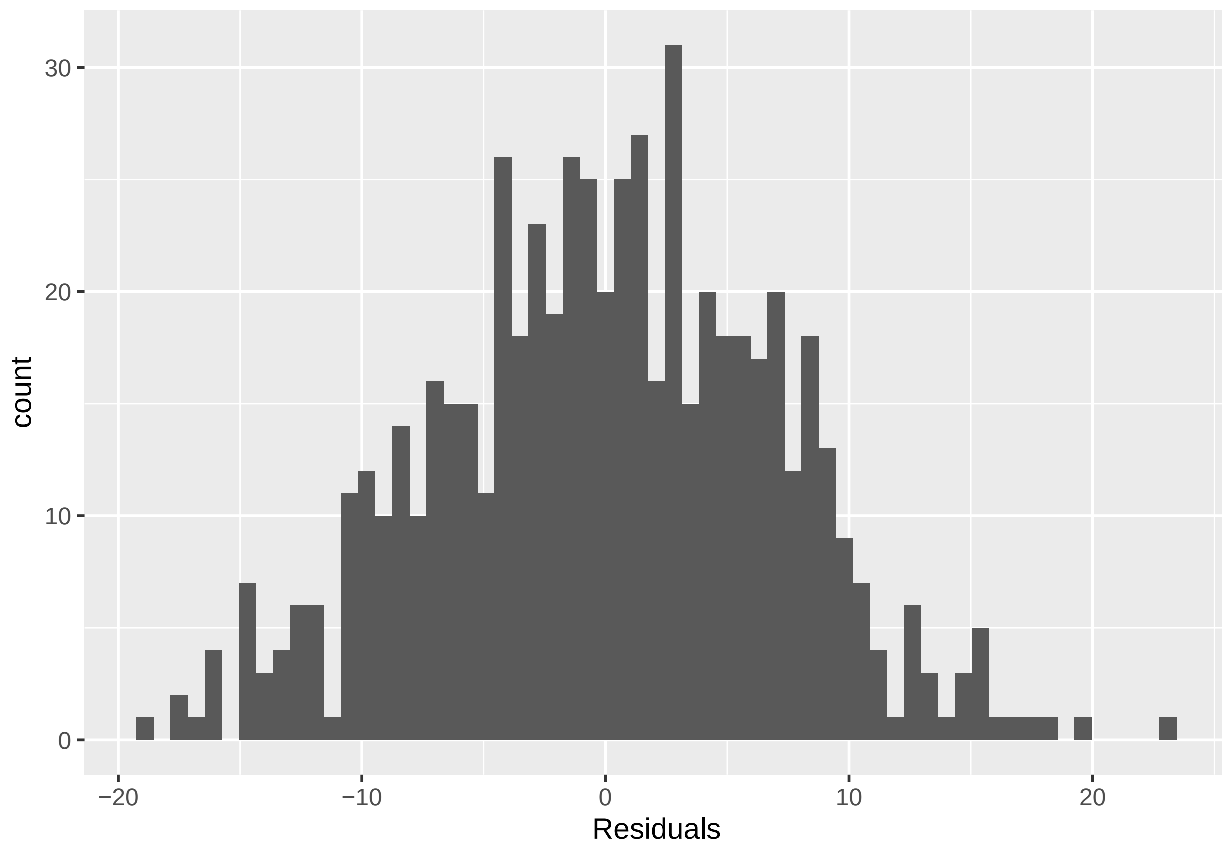
```
## `geom_smooth()` using formula 'y ~ x'
```



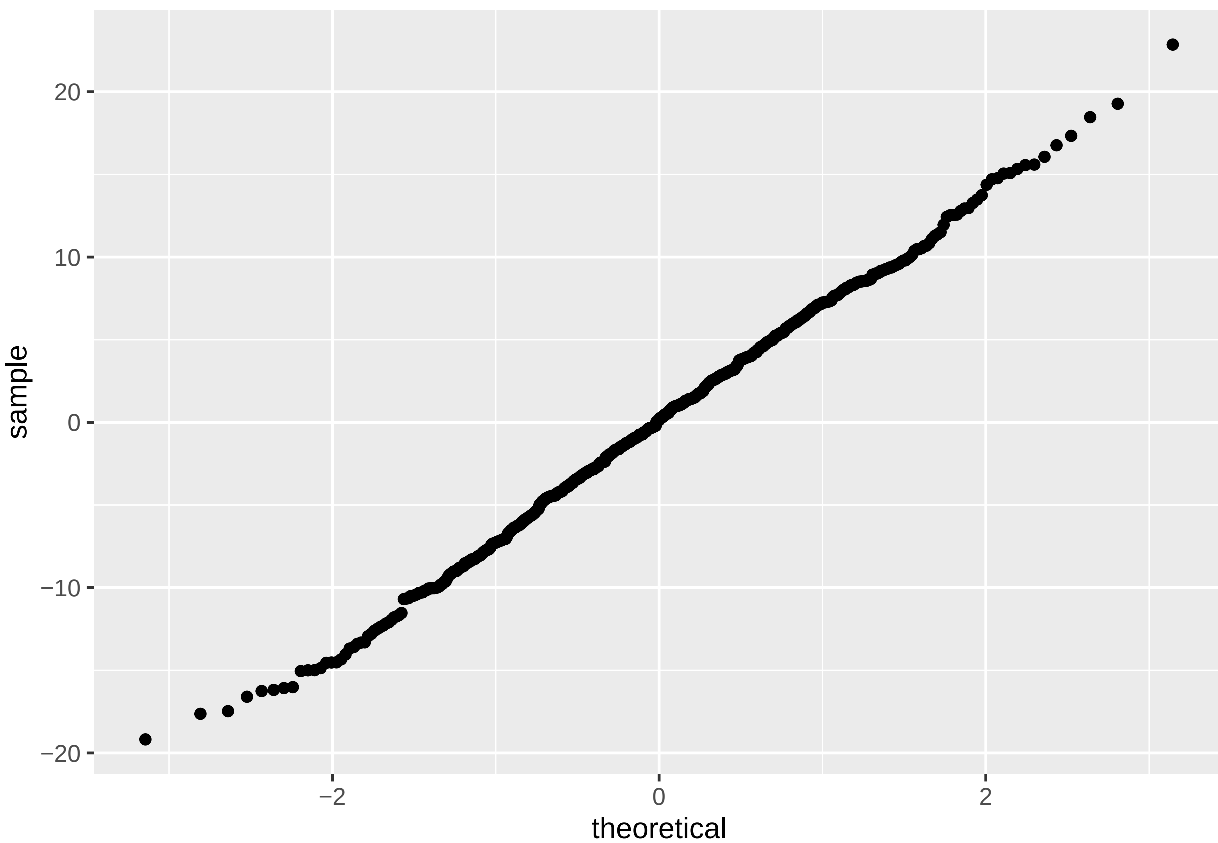
The data appear to be heavily concentrated between 0 and 125 000 votes.

2. Nearly normal residuals.

```
# histogram of residuals
ggplot(data = m_final, aes(x = .resid)) +
  geom_histogram(binwidth = 0.7) +
  xlab("Residuals")
```



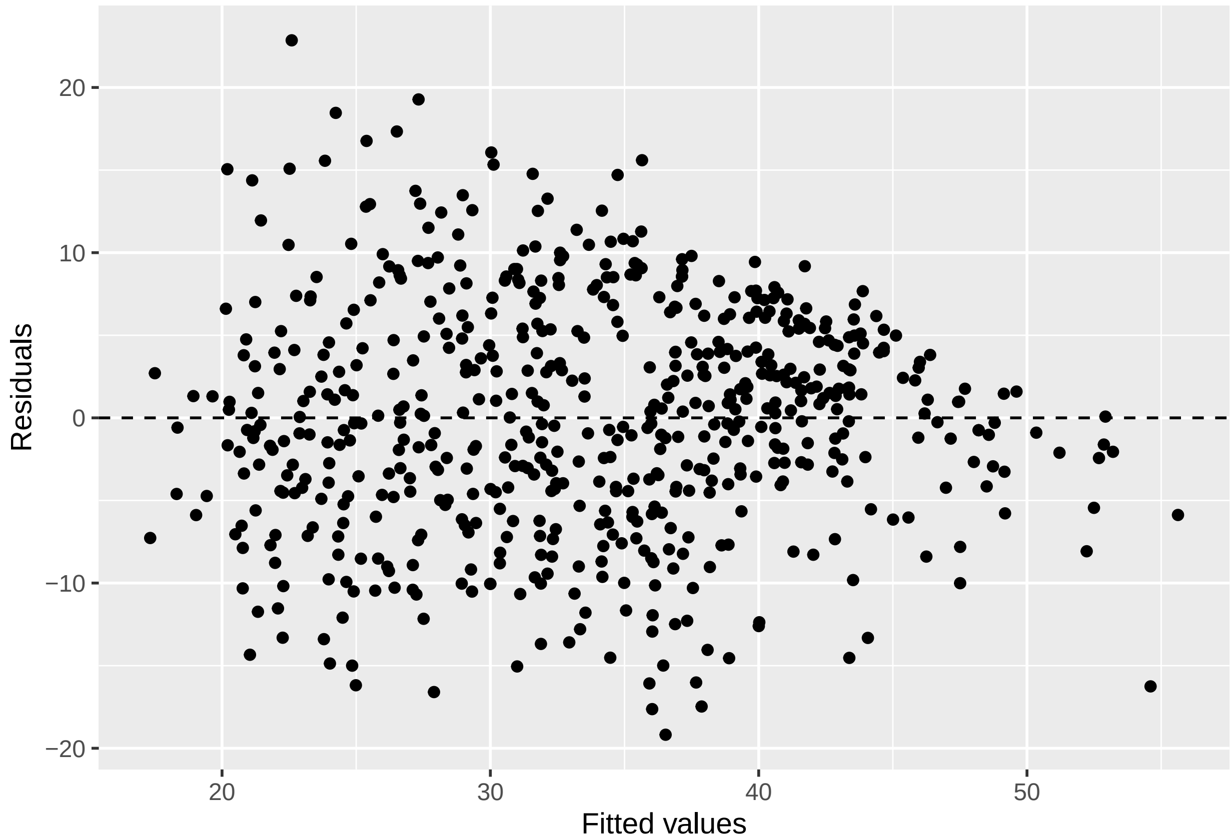

```
# normal probability plot of residuals
ggplot(data = m_final, aes(sample = .resid)) +
  stat_qq()
```



The residuals appear to be normally distributed and centered at 0.

3. Constant variability

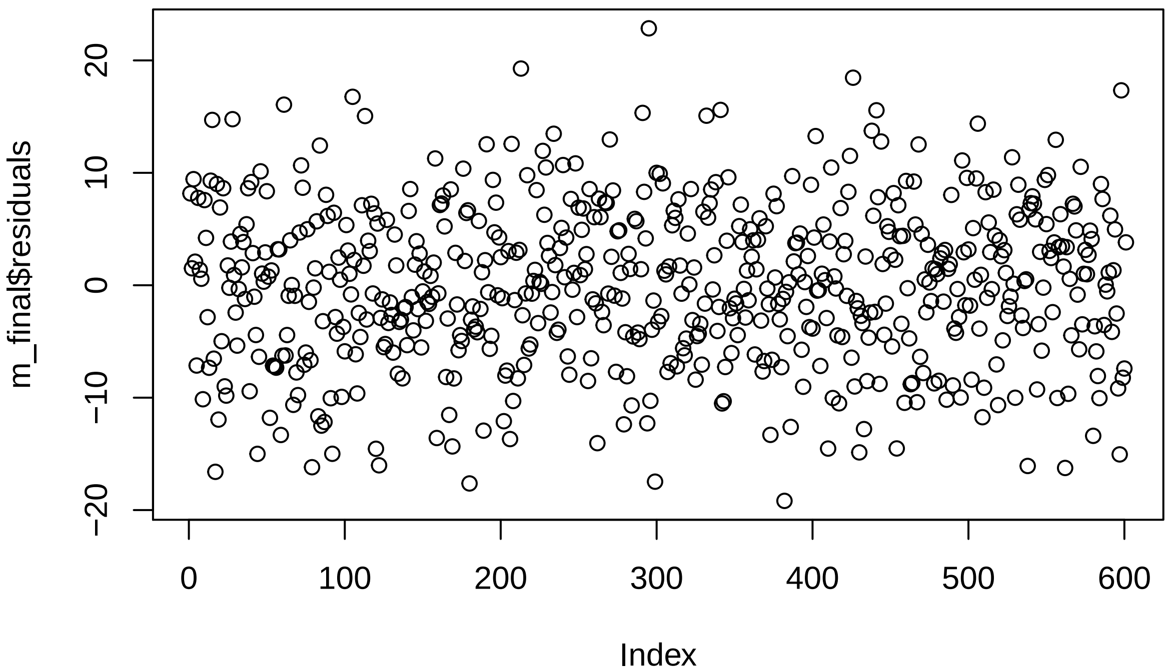
```
ggplot(data = m_final, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```



The variability around the 0 line seems to be roughly constant.

4. Independence of residuals

```
plot(m_final$residuals)
```



There appears to be no time structure in `critics_score` data collection process.

In general, the conditions for the MLR model are satisfied.

Interpretation of model coefficients

The two numerical explanatory variables, `imdb_num_votes` and `critics_score`, have a positive linear relationship with the response variable. Higher value of each of these variables, all other independent variables held constant, increases the popularity of a movie.

In case of the third explanatory variable, `genre`, the reference level is *Action & Adventure*. It means that all other independent variables held constant, *Action & Adventure* adds nothing to the response variable while other categories may increase or decrease the popularity score.

Conclusion

In terms of methodology, the model seems to meet all the criteria for Adjusted R-squared forward selection and model diagnostics.

According to the model, movie popularity depends first of all on `critics_score`, `imdb_num_votes` and `genre` variables.

At the same time a modest Adjusted R-squared of 0.5434 means that slightly more than 45% of the popularity is explained by other factors.

Part 5: Prediction

Movie chosen: *The Do-Over* (2016)

Prediction

```
newdata = data.frame(critics_score = 10, imdb_num_votes = 36697, genre="Comedy")
predict(m_final, newdata, interval="predict", level = 0.95)
```

```
##        fit     lwr      upr
## 1 22.09933 8.08065 36.11802
```

Actual popularity

```
# Rating on IMDB
imdb_r <- 5.7

# Audience score on Rotten Tomatoes
rt_r <- 40

# Actual popularity score
pop_actual <- (imdb_r + rt_r)/2
pop_actual
```

```
## [1] 22.85
```

Conclusion

A Comedy with 10% critics score and 36,697 votes on IMDB is expected to have a popularity score between 8.08065 and 36.11802, being 22.09933 the expected value.

With the actual popularity score of 22.85, 22.09933 is almost a perfect fit.

Sources

`critics_score`: https://www.rottentomatoes.com/m/the_do_over_2016

`imdb_num_votes`: https://www.imdb.com/title/tt4769836/

Part 6: Conclusion

Based on the prediction results we can conclude that the model is quite accurate at predicting films popularity.

However, if we wanted to conduct an experiment to establish a causal relationship, we wouldn't be able to do so as the numerical explanatory variables are out of our control. This can be considered as the most significant shortcoming of the developed model.

Future research ideas may include text analysis on how specific directors and cast affect the movie popularity.