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 is generally generalizable to all film because it is a properly randomised sample.

Part 2: Research question

What factors are associated with the Critics score on Rotten Tomatoes (critics_score)?

Part 3: Exploratory data analysis

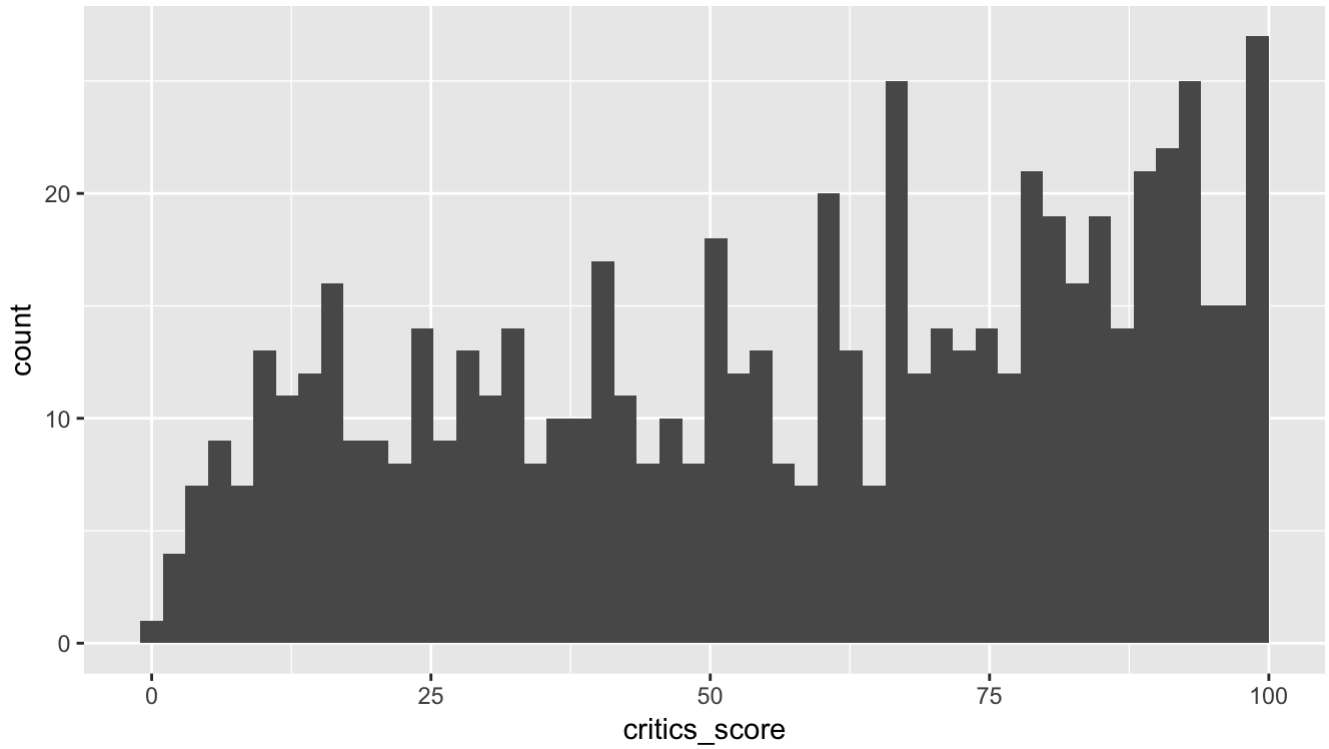
This is the summary statistics of the critics score on Rotten tomatoes.

```
summary(movies$critics_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   33.00   61.00   57.69   83.00  100.00
```

This is the histogram of the critics score

```
ggplot(data=movies, aes(x=critics_score)) + geom_histogram(bins = 50)
```



Part 4: Modeling

The following varibles will be included: title_type, genre, runtime, mpaa_rating thtr_rel_month, best_actor_win, best_actress win, best_dir_win, best_pic_win, best_pic_nom, top200_box, imdb_rating, imdb_num_votes

The following are the excluded variables with reasons

1. studio & director: There are too many studio or director, so it is difficult to analyse in a linear regression model
2. Year and date: These 2 variables are unlikely predictors logically
3. Rotten potatoes critics category: it is obviously associated with the critics score, so it is excluded
4. link: it is not a informative data for analysis
5. audience_score: it has high collinearity with imdb_rating, see below:

```
cor(movies$audience_score, movies$imdb_rating)
```

```
## [1] 0.8648652
```

For the model selection, backward selection with p-values is chosen.

Full model:

```
lm1 <- lm(data=movies, critics_score ~ title_type + genre + runtime + mpaa_rating + thtr_rel_mon
  th +  best_actor_win+ best_actress_win + best_dir_win+ best_pic_win+ best_pic_nom+ top200_box+
  imdb_rating+ imdb_num_votes )
summary(lm1)
```

```
##
## Call:
## lm(formula = critics_score ~ title_type + genre + runtime + mpaa_rating +
##     thtr_rel_month + best_actor_win + best_actress_win + best_dir_win +
##     best_pic_win + best_pic_nom + top200_box + imdb_rating +
##     imdb_num_votes, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.075 -11.781   1.568  12.394  46.974
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -4.082e+01  1.063e+01  -3.842 0.000135 ***
## title_typeFeature Film        -1.637e+01  6.651e+00  -2.462 0.014098 *
## title_typeTV Movie            -1.062e+01  1.048e+01  -1.014 0.310997
## genreAnimation                 5.936e+00  7.014e+00   0.846 0.397692
## genreArt House & International -3.362e+00  5.462e+00  -0.616 0.538386
## genreComedy                    4.158e+00  3.005e+00   1.384 0.166941
## genreDocumentary              -6.416e+00  7.160e+00  -0.896 0.370567
## genreDrama                     8.132e+00  2.635e+00   3.086 0.002121 **
## genreHorror                    5.494e+00  4.463e+00   1.231 0.218787
## genreMusical & Performing Arts  4.682e+00  6.147e+00   0.762 0.446599
## genreMystery & Suspense        4.882e+00  3.366e+00   1.450 0.147509
## genreOther                     9.700e+00  5.092e+00   1.905 0.057261 .
## genreScience Fiction & Fantasy  1.105e+01  6.365e+00   1.736 0.083115 .
## runtime                       -4.262e-02  4.464e-02  -0.955 0.340169
## mpaa_ratingNC-17               1.453e+01  1.356e+01   1.071 0.284408
## mpaa_ratingPG                 -3.783e+00  4.946e+00  -0.765 0.444716
## mpaa_ratingPG-13              -9.363e+00  5.107e+00  -1.834 0.067193 .
## mpaa_ratingR                  -5.855e+00  4.928e+00  -1.188 0.235219
## mpaa_ratingUnrated             2.851e+00  5.647e+00   0.505 0.613776
## thtr_rel_month                -1.929e-01  2.065e-01  -0.934 0.350601
## best_actor_winyes             -9.944e-02  2.131e+00  -0.047 0.962790
## best_actress_winyes            4.099e-01  2.356e+00   0.174 0.861953
## best_dir_winyes                5.885e+00  3.079e+00   1.911 0.056465 .
## best_pic_winyes               -5.902e-01  8.293e+00  -0.071 0.943285
## best_pic_nomyes                7.986e+00  4.710e+00   1.695 0.090482 .
## top200_boxyes                  8.521e+00  5.007e+00   1.702 0.089333 .
## imdb_rating                    1.834e+01  8.492e-01  21.599  < 2e-16 ***
## imdb_num_votes                -6.671e-06  7.982e-06  -0.836 0.403590
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.84 on 622 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.6223, Adjusted R-squared:  0.6059
## F-statistic: 37.95 on 27 and 622 DF,  p-value: < 2.2e-16
```

best_actor_win will be dropped due to highest p value, i.e. 0.962790

```
lm2 <- lm(data=movies, critics_score ~ title_type + genre + runtime + mpaa_rating + thtr_rel_mon
  th + best_actress_win + best_dir_win+ best_pic_win+ best_pic_nom+ top200_box+ imdb_rating+ imd
  b_num_votes )
summary(lm2)
```

```
##
## Call:
## lm(formula = critics_score ~ title_type + genre + runtime + mpaa_rating +
##     thtr_rel_month + best_actress_win + best_dir_win + best_pic_win +
##     best_pic_nom + top200_box + imdb_rating + imdb_num_votes,
##     data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.065 -11.801   1.545  12.408  46.985
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -4.079e+01  1.059e+01  -3.852 0.000129 ***
## title_typeFeature Film        -1.638e+01  6.645e+00  -2.464 0.013991 *
## title_typeTV Movie            -1.062e+01  1.047e+01  -1.014 0.310882
## genreAnimation                 5.929e+00  7.007e+00   0.846 0.397757
## genreArt House & International -3.354e+00  5.455e+00  -0.615 0.538839
## genreComedy                    4.157e+00  3.002e+00   1.385 0.166662
## genreDocumentary              -6.418e+00  7.154e+00  -0.897 0.369974
## genreDrama                     8.128e+00  2.632e+00   3.088 0.002104 **
## genreHorror                    5.499e+00  4.458e+00   1.233 0.217908
## genreMusical & Performing Arts  4.686e+00  6.142e+00   0.763 0.445752
## genreMystery & Suspense        4.868e+00  3.350e+00   1.453 0.146708
## genreOther                     9.698e+00  5.088e+00   1.906 0.057112 .
## genreScience Fiction & Fantasy  1.106e+01  6.357e+00   1.740 0.082421 .
## runtime                       -4.299e-02  4.388e-02  -0.980 0.327648
## mpaa_ratingNC-17               1.449e+01  1.352e+01   1.071 0.284383
## mpaa_ratingPG                 -3.789e+00  4.940e+00  -0.767 0.443369
## mpaa_ratingPG-13              -9.365e+00  5.102e+00  -1.835 0.066915 .
## mpaa_ratingR                  -5.856e+00  4.924e+00  -1.189 0.234767
## mpaa_ratingUnrated             2.853e+00  5.642e+00   0.506 0.613204
## thtr_rel_month                -1.932e-01  2.063e-01  -0.936 0.349382
## best_actress_winyes            4.037e-01  2.351e+00   0.172 0.863701
## best_dir_winyes                5.878e+00  3.074e+00   1.912 0.056280 .
## best_pic_winyes               -5.624e-01  8.265e+00  -0.068 0.945764
## best_pic_nomyes                7.961e+00  4.677e+00   1.702 0.089220 .
## top200_boxyes                  8.515e+00  5.002e+00   1.702 0.089193 .
## imdb_rating                    1.834e+01  8.485e-01  21.616  < 2e-16 ***
## imdb_num_votes                -6.658e-06  7.970e-06  -0.835 0.403857
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.82 on 623 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.6223, Adjusted R-squared:  0.6065
## F-statistic: 39.47 on 26 and 623 DF,  p-value: < 2.2e-16
```

Then, best_pic_win will be dropped due to highest p value, i.e. 0.945764

```
lm3 <- lm(data=movies, critics_score ~ title_type + genre + runtime + mpaa_rating + thtr_rel_mon
  th + best_actress_win + best_dir_win+ best_pic_nom+ top200_box+ imdb_rating+ imdb_num_votes )
summary(lm3)
```

```
##
## Call:
## lm(formula = critics_score ~ title_type + genre + runtime + mpaa_rating +
##     thtr_rel_month + best_actress_win + best_dir_win + best_pic_nom +
##     top200_box + imdb_rating + imdb_num_votes, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.071 -11.808   1.544  12.411  46.978
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -4.081e+01  1.057e+01  -3.859 0.000125 ***
## title_typeFeature Film        -1.637e+01  6.639e+00  -2.466 0.013940 *
## title_typeTV Movie            -1.062e+01  1.046e+01  -1.015 0.310432
## genreAnimation                 5.929e+00  7.001e+00   0.847 0.397389
## genreArt House & International -3.366e+00  5.448e+00  -0.618 0.536925
## genreComedy                    4.148e+00  2.997e+00   1.384 0.166833
## genreDocumentary              -6.427e+00  7.147e+00  -0.899 0.368870
## genreDrama                     8.124e+00  2.629e+00   3.090 0.002091 **
## genreHorror                    5.494e+00  4.454e+00   1.233 0.217901
## genreMusical & Performing Arts  4.679e+00  6.136e+00   0.763 0.445964
## genreMystery & Suspense        4.864e+00  3.347e+00   1.453 0.146649
## genreOther                     9.719e+00  5.075e+00   1.915 0.055922 .
## genreScience Fiction & Fantasy  1.106e+01  6.351e+00   1.742 0.081993 .
## runtime                       -4.303e-02  4.384e-02  -0.981 0.326758
## mpaa_ratingNC-17               1.448e+01  1.351e+01   1.072 0.284025
## mpaa_ratingPG                 -3.788e+00  4.936e+00  -0.767 0.443188
## mpaa_ratingPG-13              -9.352e+00  5.094e+00  -1.836 0.066883 .
## mpaa_ratingR                  -5.849e+00  4.919e+00  -1.189 0.234829
## mpaa_ratingUnrated             2.855e+00  5.637e+00   0.506 0.612702
## thtr_rel_month                -1.924e-01  2.058e-01  -0.935 0.350206
## best_actress_winyes            3.964e-01  2.346e+00   0.169 0.865890
## best_dir_winyes                5.820e+00  2.951e+00   1.972 0.049024 *
## best_pic_nomyes                7.834e+00  4.282e+00   1.830 0.067797 .
## top200_boxyes                  8.516e+00  4.998e+00   1.704 0.088908 .
## imdb_rating                    1.835e+01  8.459e-01  21.688  < 2e-16 ***
## imdb_num_votes                -6.756e-06  7.832e-06  -0.863 0.388645
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.81 on 624 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.6223, Adjusted R-squared:  0.6071
## F-statistic: 41.12 on 25 and 624 DF,  p-value: < 2.2e-16
```

Next, best_actress_win due to highest p-value 0.86589.

```
lm4 <- lm(data=movies, critics_score ~ title_type + genre + runtime + mpaa_rating + thtr_rel_mon
  th + best_dir_win+ best_pic_nom+ top200_box+ imdb_rating+ imdb_num_votes)
summary(lm4)
```

```
##
## Call:
## lm(formula = critics_score ~ title_type + genre + runtime + mpaa_rating +
##     thtr_rel_month + best_dir_win + best_pic_nom + top200_box +
##     imdb_rating + imdb_num_votes, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.120 -11.860   1.517  12.410  46.985
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -4.091e+01  1.055e+01  -3.878 0.000117 ***
## title_typeFeature Film        -1.637e+01  6.634e+00  -2.467 0.013884 *
## title_typeTV Movie            -1.057e+01  1.045e+01  -1.012 0.312054
## genreAnimation                 5.992e+00  6.986e+00   0.858 0.391393
## genreArt House & International -3.323e+00  5.438e+00  -0.611 0.541334
## genreComedy                    4.197e+00  2.980e+00   1.408 0.159539
## genreDocumentary              -6.396e+00  7.139e+00  -0.896 0.370621
## genreDrama                     8.181e+00  2.605e+00   3.140 0.001770 **
## genreHorror                    5.516e+00  4.449e+00   1.240 0.215517
## genreMusical & Performing Arts  4.681e+00  6.131e+00   0.763 0.445483
## genreMystery & Suspense        4.930e+00  3.321e+00   1.484 0.138231
## genreOther                     9.748e+00  5.068e+00   1.924 0.054857 .
## genreScience Fiction & Fantasy  1.107e+01  6.346e+00   1.744 0.081698 .
## runtime                       -4.216e-02  4.350e-02  -0.969 0.332909
## mpaa_ratingNC-17               1.444e+01  1.350e+01   1.070 0.285102
## mpaa_ratingPG                 -3.779e+00  4.932e+00  -0.766 0.443875
## mpaa_ratingPG-13              -9.344e+00  5.090e+00  -1.836 0.066884 .
## mpaa_ratingR                  -5.855e+00  4.915e+00  -1.191 0.233971
## mpaa_ratingUnrated             2.844e+00  5.633e+00   0.505 0.613785
## thtr_rel_month                -1.927e-01  2.056e-01  -0.937 0.349222
## best_dir_winyes                5.829e+00  2.948e+00   1.977 0.048479 *
## best_pic_nomyes                7.932e+00  4.238e+00   1.872 0.061727 .
## top200_boxyes                  8.560e+00  4.987e+00   1.716 0.086583 .
## imdb_rating                    1.835e+01  8.452e-01  21.705  < 2e-16 ***
## imdb_num_votes                -6.751e-06  7.826e-06  -0.863 0.388672
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.79 on 625 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.6222, Adjusted R-squared:  0.6077
## F-statistic:  42.9 on 24 and 625 DF,  p-value: < 2.2e-16
```

imdb_num_votes will be dropped due to highest p value 0.388672 P.S. Some p values in a certain level of variables are higher than 0.388672, but the other levels of the same variable had p value lower than 0.388672, so that variable is kept (e.g. mpaa_rating)

```
lm5 <- lm(data=movies, critics_score ~ title_type + genre + runtime + mpaa_rating + thtr_rel_mon
  th + best_dir_win+ best_pic_nom+ top200_box+ imdb_rating)
summary(lm5)
```

```
##
## Call:
## lm(formula = critics_score ~ title_type + genre + runtime + mpaa_rating +
##     thtr_rel_month + best_dir_win + best_pic_nom + top200_box +
##     imdb_rating, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.676 -11.964   1.133  12.434  47.147
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -38.70435   10.23402  -3.782 0.000171 ***
## title_typeFeature Film        -16.63867    6.62498  -2.512 0.012272 *
## title_typeTV Movie            -10.73699   10.44237  -1.028 0.304246
## genreAnimation                  5.71765    6.97706   0.819 0.412817
## genreArt House & International  -2.79237    5.40162  -0.517 0.605373
## genreComedy                     4.31271    2.97669   1.449 0.147887
## genreDocumentary               -5.98813    7.12212  -0.841 0.400793
## genreDrama                      8.54985    2.56949   3.327 0.000928 ***
## genreHorror                     5.69641    4.44293   1.282 0.200272
## genreMusical & Performing Arts   5.25699    6.09336   0.863 0.388610
## genreMystery & Suspense         5.12798    3.31282   1.548 0.122147
## genreOther                      9.71683    5.06646   1.918 0.055582 .
## genreScience Fiction & Fantasy  10.96882    6.34370   1.729 0.084286 .
## runtime                        -0.04830    0.04291  -1.126 0.260713
## mpaa_ratingNC-17               14.35079   13.49200   1.064 0.287897
## mpaa_ratingPG                  -3.86834    4.93001  -0.785 0.432954
## mpaa_ratingPG-13               -9.75929    5.06642  -1.926 0.054523 .
## mpaa_ratingR                   -6.17822    4.89940  -1.261 0.207773
## mpaa_ratingUnrated              2.78353    5.63103   0.494 0.621253
## thtr_rel_month                 -0.19166    0.20560  -0.932 0.351604
## best_dir_winyes                 5.66898    2.94185   1.927 0.054430 .
## best_pic_nomyes                 7.17402    4.14525   1.731 0.084005 .
## top200_boxyes                   7.52128    4.83870   1.554 0.120594
## imdb_rating                    18.09417    0.79323  22.811  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.79 on 626 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.6218, Adjusted R-squared:  0.6079
## F-statistic: 44.75 on 23 and 626 DF,  p-value: < 2.2e-16
```

Now, thtr_rel_month is dropped with p-value of 0.351604

```
lm6 <- lm(data=movies, critics_score ~ title_type + genre + runtime + mpaa_rating + best_dir_win
  + best_pic_nom+ top200_box+ imdb_rating)
summary(lm6)
```

```
##
## Call:
## lm(formula = critics_score ~ title_type + genre + runtime + mpaa_rating +
##     best_dir_win + best_pic_nom + top200_box + imdb_rating, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.905 -11.685   1.248  12.210  47.043
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -39.21622   10.21821  -3.838 0.000137 ***
## title_typeFeature Film        -16.59985    6.62415  -2.506 0.012464 *
## title_typeTV Movie            -10.44077   10.43644  -1.000 0.317496
## genreAnimation                  5.44982    6.97041   0.782 0.434597
## genreArt House & International -2.80207    5.40105  -0.519 0.604083
## genreComedy                     4.19575    2.97373   1.411 0.158759
## genreDocumentary               -5.98294    7.12138  -0.840 0.401151
## genreDrama                      8.57646    2.56907   3.338 0.000893 ***
## genreHorror                     5.67449    4.44240   1.277 0.201953
## genreMusical & Performing Arts  5.16628    6.09194   0.848 0.396733
## genreMystery & Suspense         5.27784    3.30857   1.595 0.111170
## genreOther                      9.95916    5.05926   1.968 0.049451 *
## genreScience Fiction & Fantasy 11.02595    6.34274   1.738 0.082639 .
## runtime                        -0.05589    0.04213  -1.327 0.185092
## mpaa_ratingNC-17               14.57344   13.48847   1.080 0.280363
## mpaa_ratingPG                  -3.94087    4.92888  -0.800 0.424276
## mpaa_ratingPG-13               -9.67372    5.06505  -1.910 0.056603 .
## mpaa_ratingR                   -6.24662    4.89834  -1.275 0.202692
## mpaa_ratingUnrated              2.82725    5.63024   0.502 0.615736
## best_dir_winyes                 5.64324    2.94141   1.919 0.055496 .
## best_pic_nomyes                 6.72896    4.11723   1.634 0.102689
## top200_boxyes                   7.28899    4.83177   1.509 0.131916
## imdb_rating                    18.09775    0.79313  22.818  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.79 on 627 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.6213, Adjusted R-squared:  0.608
## F-statistic: 46.75 on 22 and 627 DF,  p-value: < 2.2e-16
```

runtime is then dropped due to highest p-value of 0.185092

```
lm7 <- lm(data=movies, critics_score ~ title_type + genre + mpaa_rating + best_dir_win+ best_pic
  _nom+ top200_box+ imdb_rating)
summary(lm7)
```

```
##
## Call:
## lm(formula = critics_score ~ title_type + genre + mpaa_rating +
##     best_dir_win + best_pic_nom + top200_box + imdb_rating, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.314 -12.060   1.373  12.326  47.359
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -42.8250     9.8577  -4.344 1.63e-05 ***
## title_typeFeature Film         -16.7391     6.6221  -2.528  0.01172 *
## title_typeTV Movie             -10.2305    10.4340  -0.980  0.32722
## genreAnimation                   5.7324     6.9666   0.823  0.41091
## genreArt House & International  -2.5444     5.3977  -0.471  0.63753
## genreComedy                      4.5716     2.9603   1.544  0.12302
## genreDocumentary                -5.4222     7.1043  -0.763  0.44561
## genreDrama                       8.4179     2.5661   3.280  0.00109 **
## genreHorror                      6.2507     4.4218   1.414  0.15797
## genreMusical & Performing Arts   4.8669     6.0865   0.800  0.42423
## genreMystery & Suspense          5.0873     3.3053   1.539  0.12427
## genreOther                       9.7905     5.0574   1.936  0.05333 .
## genreScience Fiction & Fantasy  11.1232     6.3421   1.754  0.07994 .
## mpaa_ratingNC-17                14.2463    13.4859   1.056  0.29120
## mpaa_ratingPG                   -4.5887     4.9053  -0.935  0.34992
## mpaa_ratingPG-13               -10.7517     5.0011  -2.150  0.03195 *
## mpaa_ratingR                    -7.0039     4.8663  -1.439  0.15057
## mpaa_ratingUnrated               1.6916     5.5709   0.304  0.76150
## best_dir_winyes                  4.9219     2.8912   1.702  0.08918 .
## best_pic_nomyes                  5.7989     4.0578   1.429  0.15348
## top200_boxyes                    6.5921     4.8035   1.372  0.17044
## imdb_rating                     17.8955     0.7776  23.013  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.79 on 629 degrees of freedom
## Multiple R-squared:  0.6205, Adjusted R-squared:  0.6078
## F-statistic: 48.96 on 21 and 629 DF,  p-value: < 2.2e-16
```

Now top200_box is dropped with highest p-value of 0.17044.

```
lm8 <- lm(data=movies, critics_score ~ title_type + genre + mpaa_rating + best_dir_win+ best_pic
  _nom+ imdb_rating)
summary(lm8)
```

```
##
## Call:
## lm(formula = critics_score ~ title_type + genre + mpaa_rating +
##     best_dir_win + best_pic_nom + imdb_rating, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.364 -11.986   1.481  12.197  46.797
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -42.6572     9.8638  -4.325 1.78e-05 ***
## title_typeFeature Film         -16.6023     6.6260  -2.506  0.01247 *
## title_typeTV Movie             -10.1134    10.4410  -0.969  0.33310
## genreAnimation                   4.8419     6.9411   0.698  0.48571
## genreArt House & International  -2.9974     5.3913  -0.556  0.57843
## genreComedy                      4.2208     2.9513   1.430  0.15317
## genreDocumentary                -5.9545     7.0986  -0.839  0.40189
## genreDrama                       8.0368     2.5528   3.148  0.00172 **
## genreHorror                      5.9239     4.4185   1.341  0.18050
## genreMusical & Performing Arts   4.3289     6.0781   0.712  0.47660
## genreMystery & Suspense          4.7410     3.2980   1.438  0.15105
## genreOther                       9.6201     5.0594   1.901  0.05770 .
## genreScience Fiction & Fantasy  11.3692     6.3440   1.792  0.07359 .
## mpaa_ratingNC-17                13.4781    13.4837   1.000  0.31789
## mpaa_ratingPG                   -5.0623     4.8966  -1.034  0.30160
## mpaa_ratingPG-13               -11.3303     4.9868  -2.272  0.02342 *
## mpaa_ratingR                    -7.7385     4.8401  -1.599  0.11036
## mpaa_ratingUnrated               1.0249     5.5535   0.185  0.85365
## best_dir_winyes                  4.9668     2.8931   1.717  0.08651 .
## best_pic_nomyes                  6.0945     4.0549   1.503  0.13335
## imdb_rating                     18.0216     0.7727  23.322  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.8 on 630 degrees of freedom
## Multiple R-squared:  0.6193, Adjusted R-squared:  0.6072
## F-statistic: 51.25 on 20 and 630 DF,  p-value: < 2.2e-16
```

best_pic_nom will be dropped due to high p-value 0.13335.

```
lm9 <- lm(data=movies, critics_score ~ title_type + genre + mpaa_rating + best_dir_win+ imdb_rat
  ing)
summary(lm9)
```

```
##
## Call:
## lm(formula = critics_score ~ title_type + genre + mpaa_rating +
##     best_dir_win + imdb_rating, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.572 -11.933   1.965  12.253  46.829
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -44.5080     9.7964  -4.543 6.64e-06 ***
## title_typeFeature Film        -16.4530     6.6319  -2.481  0.01336 *
## title_typeTV Movie            -10.1032    10.4514  -0.967  0.33407
## genreAnimation                  4.9969     6.9473   0.719  0.47225
## genreArt House & International  -3.0452     5.3966  -0.564  0.57276
## genreComedy                     4.3431     2.9531   1.471  0.14187
## genreDocumentary               -6.0839     7.1052  -0.856  0.39218
## genreDrama                      8.2172     2.5525   3.219  0.00135 **
## genreHorror                     6.0562     4.4220   1.370  0.17132
## genreMusical & Performing Arts   4.0840     6.0820   0.671  0.50215
## genreMystery & Suspense         4.8383     3.3006   1.466  0.14318
## genreOther                     10.2336     5.0480   2.027  0.04305 *
## genreScience Fiction & Fantasy  11.4264     6.3502   1.799  0.07244 .
## mpaa_ratingNC-17               13.3271    13.4968   0.987  0.32381
## mpaa_ratingPG                  -4.8404     4.8993  -0.988  0.32354
## mpaa_ratingPG-13              -11.0537     4.9884  -2.216  0.02705 *
## mpaa_ratingR                   -7.6284     4.8444  -1.575  0.11583
## mpaa_ratingUnrated              1.0248     5.5591   0.184  0.85381
## best_dir_winyes                 5.4539     2.8777   1.895  0.05852 .
## imdb_rating                    18.2712     0.7554  24.187  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.82 on 631 degrees of freedom
## Multiple R-squared:  0.618,  Adjusted R-squared:  0.6064
## F-statistic: 53.72 on 19 and 631 DF,  p-value: < 2.2e-16
```

best_dir_win will be dropped due to p value 0.05852

```
lm10 <- lm(data=movies, critics_score ~ title_type + genre + mpaa_rating + imdb_rating)
summary(lm10)
```

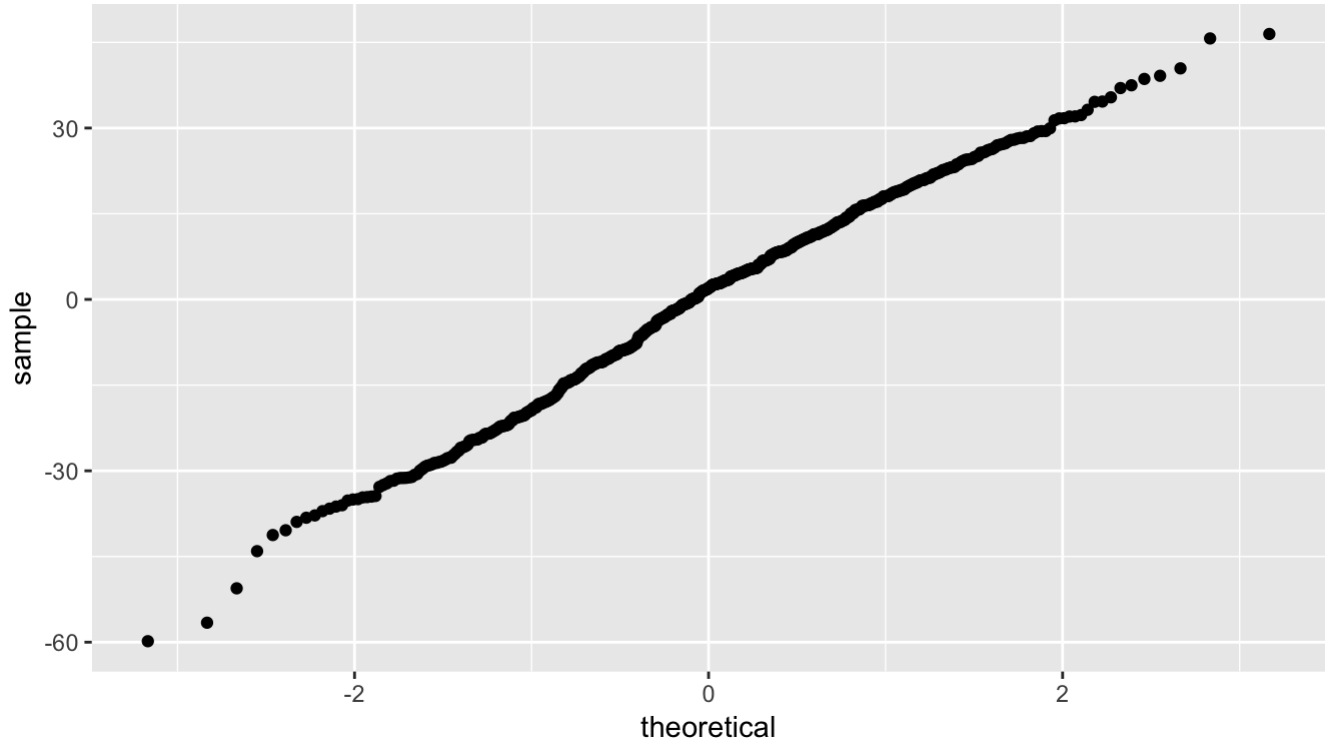
```
##
## Call:
## lm(formula = critics_score ~ title_type + genre + mpaa_rating +
##     imdb_rating, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.826 -11.936   1.996  12.241  46.465
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -46.4310     9.7637  -4.755 2.45e-06 ***
## title_typeFeature Film       -16.0108     6.6413  -2.411  0.01620 *
## title_typeTV Movie            -9.8574    10.4720  -0.941  0.34691
## genreAnimation                 4.9144     6.9614   0.706  0.48048
## genreArt House & International  -3.4931     5.4025  -0.647  0.51815
## genreComedy                    4.3011     2.9591   1.454  0.14657
## genreDocumentary              -6.2678     7.1191  -0.880  0.37896
## genreDrama                     8.1514     2.5575   3.187  0.00151 **
## genreHorror                    5.9996     4.4310   1.354  0.17622
## genreMusical & Performing Arts   4.0384     6.0944   0.663  0.50780
## genreMystery & Suspense        4.9194     3.3071   1.488  0.13737
## genreOther                    10.0674     5.0575   1.991  0.04696 *
## genreScience Fiction & Fantasy  11.7577     6.3608   1.848  0.06500 .
## mpaa_ratingNC-17              13.2717    13.5244   0.981  0.32681
## mpaa_ratingPG                 -4.3826     4.9033  -0.894  0.37177
## mpaa_ratingPG-13              -10.6681     4.9944  -2.136  0.03306 *
## mpaa_ratingR                  -7.2145     4.8494  -1.488  0.13732
## mpaa_ratingUnrated             1.2014     5.5697   0.216  0.82929
## imdb_rating                   18.5102     0.7463  24.802  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.85 on 632 degrees of freedom
## Multiple R-squared:  0.6158, Adjusted R-squared:  0.6048
## F-statistic: 56.27 on 18 and 632 DF,  p-value: < 2.2e-16
```

This is the final model, without any variables with p-value > 0.05

Model diagnostics

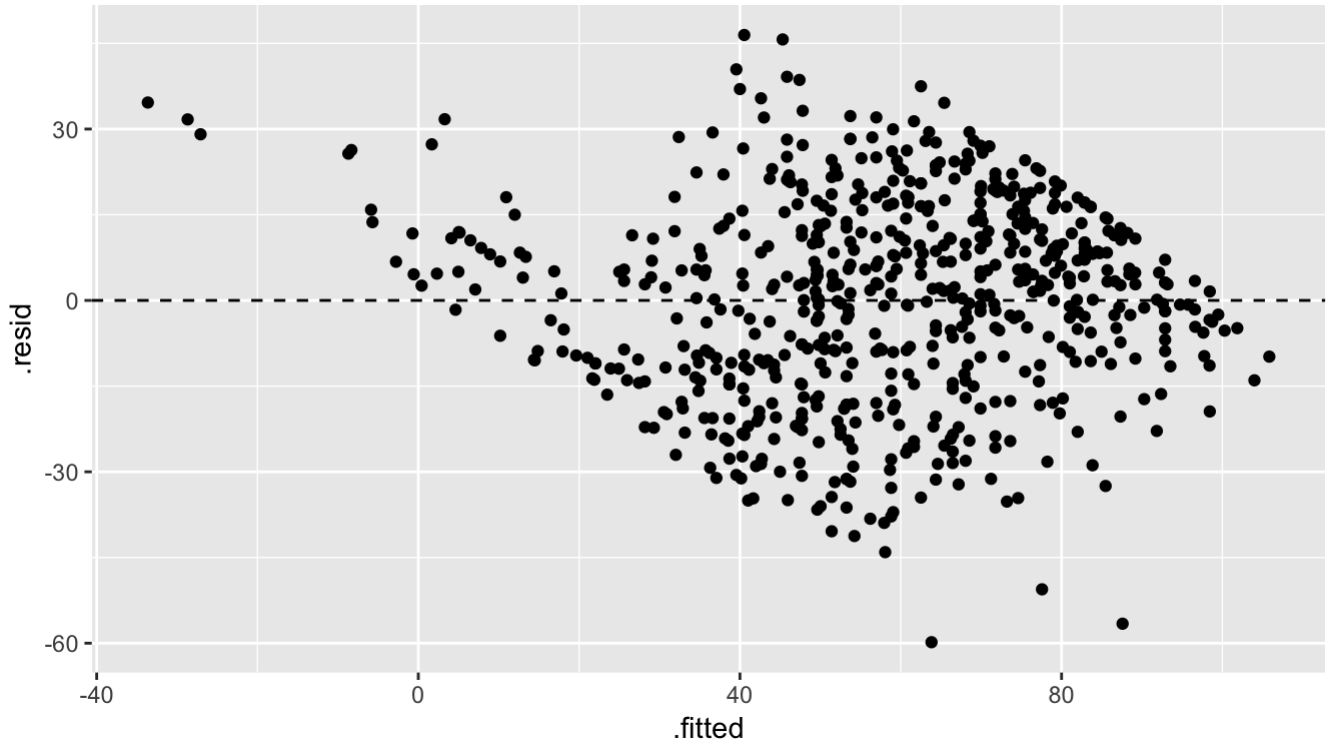
Normal probability plot: It showed a nearly straight line, so the residuals follows a nearly normal distribution

```
ggplot(data=lm10, aes(sample=.resid)) + geom_qq()
```



Residuals plot: it showed that the residuals may not be randomly scattered, but not obviously fan shape is noted

```
ggplot(data=lm10, aes(x=.fitted, y=.resid)) + geom_point() + geom_hline(yintercept = 0, linetype
  = "dashed")
```



Coefficient intepretation

Take imdb_rating as an example, for every one point increase in imdb_rating, the critics score on rotten potatoes will increased by 18.5102 score on average.

For categorical variables, for example genre, if the genre is Drama, the critics score on rotten potatoes will be increased by 8.1514 on average compared to the reference level “Action & Adventure”. Other coefficient are interpreted similarly.

Part 5: Prediction

I’ve chosen the Angry bird movie in 2016, the following links are where the information are extracted.

https://www.imdb.com/title/tt1985949/?ref_=ttls_li_tt (https://www.imdb.com/title/tt1985949/?ref_=ttls_li_tt)
https://www.rottentomatoes.com/m/the_angry_birds_movie
(https://www.rottentomatoes.com/m/the_angry_birds_movie)
https://www.upstateparent.com/story/entertainment/movies/2016/05/18/review-angry-birds-movie-what-parents-should-know/84556478/
(https://www.upstateparent.com/story/entertainment/movies/2016/05/18/review-angry-birds-movie-what-parents-should-know/84556478/)

Let’s create a newmovie data frame first

```
newmovie<- data.frame (title_type="Feature Film", genre = "Animation", mpaa_rating ="PG", imdb_r
  ating = 6.3)
```

Then predict the critics score, the predicted score is 54.70461 with range 17.2-92.2 (rounded)

```
predict(lm10, newmovie)
```

```
##        1
## 54.70461
```

```
predict(lm10, newmovie, interval="prediction", level=0.95)
```

```
##        fit      lwr      upr
## 1 54.70461 17.18054 92.22867
```

Therefore, this model predict, with 95% confience, that a features animation film with PG rating and IMDF rating of 6.3 will have a critic score on rotten potatoes in between 17.2-92.2.

Part 6: Conclusion

In conclusion, type, genre, mpaa rating, imdb rating of film are significantly associated with the critics score on rotten potatoes. However, there are no causal relationship between the response and explanatory variable.