Modeling and Prediction for Movies

Kelly Radimer

Setup

Load packages

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

Load data

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

Part 1: Data

The 651 movies in the data set were selected randomly, so our inferences should be extended to all movies in the population (movies released before 2016). There was no random assignment, so it is not appropriate to assume that there is a causal relationship between any of the variables observed.

Part 2: Research question

Since the audience's enjoyment of a movie is an important indicator of its success, our goal will be to create a model that predicts, as accurately as possible, the audience score on Rotten Tomatoes. We will consider only factors that come into play prior to the release of the movie, since we are trying to create a model that we can use to create movies that audiences will like. Specifically, we will analyze the predictive significance of type of movie, genre, runtime, mpaa rating, month of release, whether the director had won a best director Oscar, and whether one of the actors or actresses was a Best Actor/Actress winner. We need not consider production studio, since this project is being done for Paramount Pictures.

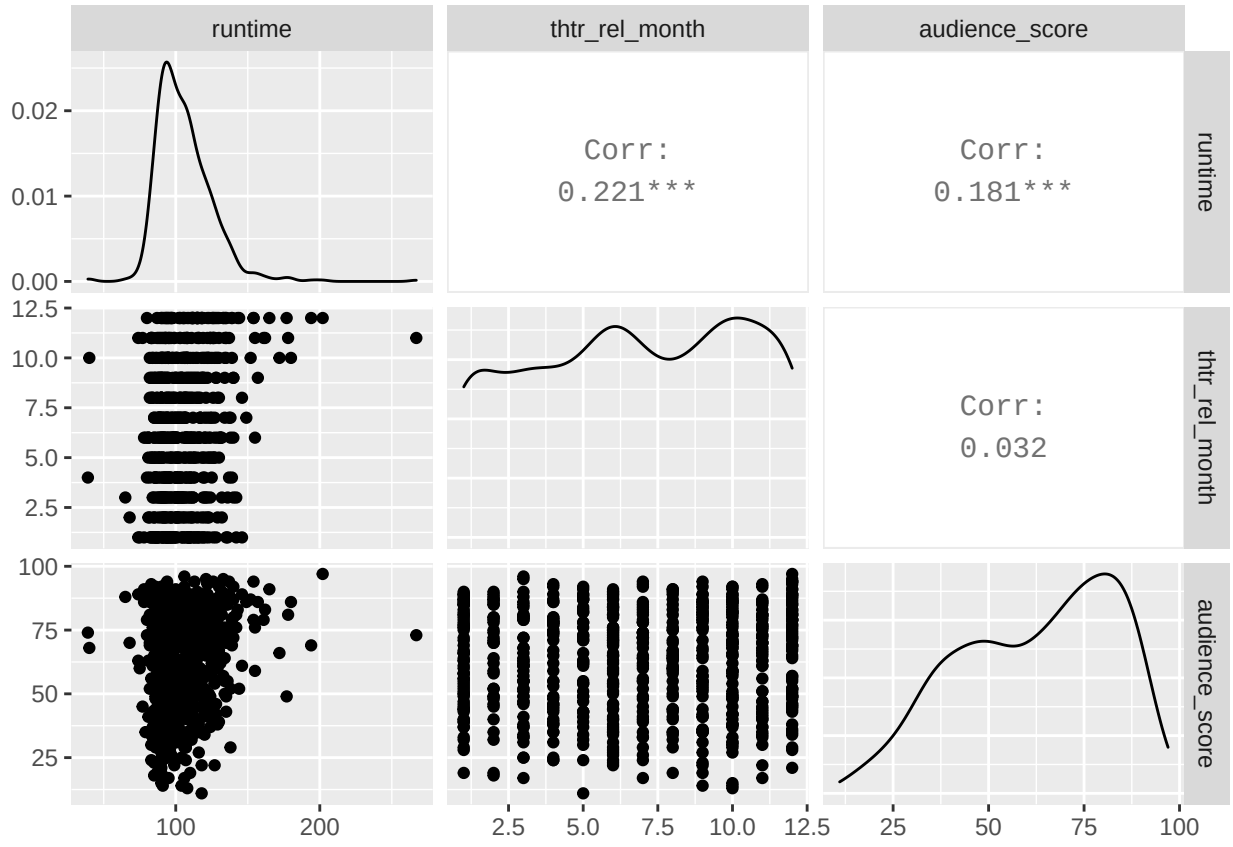
Part 3: Exploratory data analysis

Let's start by cleaning up the data. There are blanks in the runtime data that we should eliminate, since all movies have a run time and this must be an error.

```
movies <- movies %>%
  filter(!is.na(runtime))
```

Now let's have a look at the relationship between our quantitative predictors and audience scores.

```
ggpairs(movies, columns = c(4,8,18))
```




```
summary(movies$runtime)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    39.0    92.0   103.0   105.8   115.8   267.0
```

```
summary(movies$audience_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.00   46.00   65.00   62.35   80.00   97.00
```

We can see that the correlation is weak and positive between audience scores and run time (0.181). The correlation between audience scores and release month is almost nonexistent, suggesting that a viewer's feelings about a movie have little to do with the time of year that the movie came out.

We can see from the density curves that release month is nearly uniformly distributed with a peak in the middle (likely representing the summer months) and another peak near the end of the year (the Thanksgiving/Christmas time period) when more movies than usual are released.

Run times are right skewed with a small IQR. The middle 50% of movies last between 92 and 115 minutes. There are a handful of high outliers, with a maximum at 267 minutes.

Audience scores are left skewed with a median of 65, a minimum of 11 and a maximum of 97. The IQR is 34.

Now let's have a look at our categorical variables' relationship with audience scores.

```
movies %>%
  group_by(title_type) %>%
  summarise(mean=mean(audience_score))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 2
##   title_type    mean
##   <fct>        <dbl>
## 1 Documentary   83.5
## 2 Feature Film  60.5
## 3 TV Movie      56.8
```

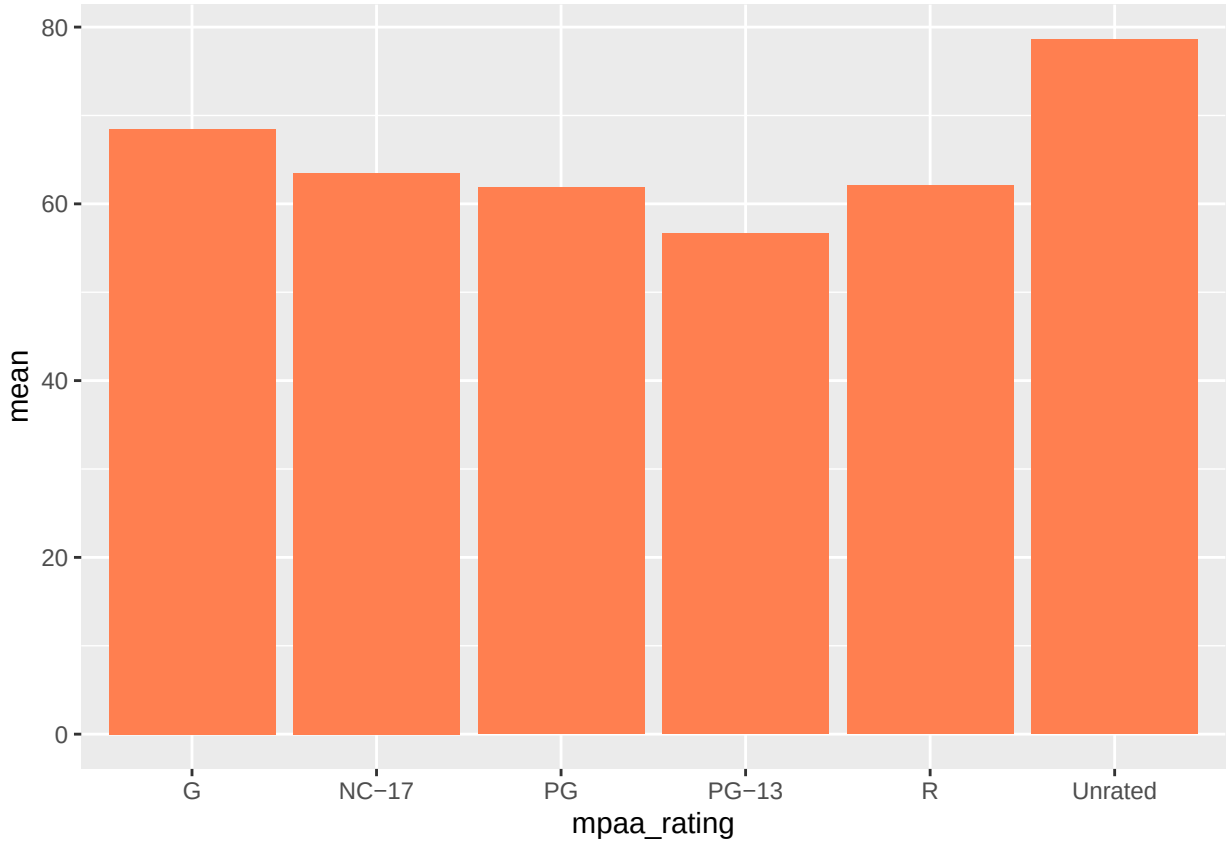
We can see that the mean audience score is highest for Documentaries, followed by Feature Films, with the lowest scores, on average, given to TV Movies.

```
audscorebyrating <- movies %>%
  group_by(mpaa_rating)%>%
  summarise(mean=mean(audience_score)) %>%
  arrange(desc(mean))

audscorebyrating
```

```
## # A tibble: 6 x 2
##   mpaa_rating  mean
##   <fct>       <dbl>
## 1 Unrated      78.6
## 2 G            68.5
## 3 NC-17        63.5
## 4 R            62.0
## 5 PG           61.8
## 6 PG-13        56.7
```

```
ggplot(data = audscorebyrating, aes(x=mpaa_rating, y=mean))+geom_col(fill="coral")
```



Unrated movies have the highest average rating, followed by G movies, NC-17 movies, and then R movies. The lowest average ratings are for PG-13 movies.

```
movies%>%
  group_by(best_actor_win)%>%
  summarise(mean=mean(audience_score))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 2
##   best_actor_win  mean
##   <fct>          <dbl>
## 1 no              62.2
## 2 yes             63.3
```

Mean ratings for films with a Best Actor Award Winner are about one point higher than those without.

```
movies%>%
  group_by(best_actress_win)%>%
  summarise(mean=mean(audience_score))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 2
```

```
##   best_actress_win  mean
##   <fct>            <dbl>
## 1 no                62.2
## 2 yes               63.9
```

Mean ratings for films with a Best Actress Award Winner are about 1.7 points higher than those without.

```
movies%>%
  group_by(genre)%>%
  summarise(mean=mean(audience_score)) %>%
  arrange(desc(mean))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 11 x 2
##    genre                      mean
##    <fct>                     <dbl>
##  1 Documentary                83.0
##  2 Musical & Performing Arts  80.2
##  3 Other                      66.7
##  4 Drama                      65.3
##  5 Art House & International  64
##  6 Animation                  62.4
##  7 Mystery & Suspense         55.9
##  8 Action & Adventure         53.8
##  9 Comedy                     52.5
## 10 Science Fiction & Fantasy  50.9
## 11 Horror                     45.8
```

The genre with the highest average rating is Documentary, followed closely by Musical and Performing Arts. The lowest average ratings go to Horror movies, followed by Sci-Fi/Fantasy and Comedies.

```
movies%>%
  group_by(best_dir_win)%>%
  summarise(mean=mean(audience_score))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 2
##   best_dir_win  mean
##   <fct>        <dbl>
## 1 no            61.8
## 2 yes           69.5
```

Movies directed by Oscar winning directors get rated about 7.7 points higher on average than movies directed by people who haven't won Best Director.

Part 4: Modeling

We will begin with a model that takes into account all 8 of our predictors of interest.

```
full_model<-lm(audience_score ~ title_type + genre + runtime + mpaa_rating + thtr_rel_month + best_acto
summary(full_model)
```

```
##
## Call:
## lm(formula = audience_score ~ title_type + genre + runtime +
##     mpaa_rating + thtr_rel_month + best_actor_win + best_actress_win +
##     best_dir_win, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.684 -12.995   1.278  13.251  39.492
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     58.05485    9.00873   6.444 2.32e-10 ***
## title_typeFeature Film         -11.83293    6.61174  -1.790  0.07399 .
## title_typeTV Movie             -21.40435   10.41533  -2.055  0.04028 *
## genreAnimation                   4.67673    6.97068   0.671  0.50252
## genreArt House & International   9.53012    5.39195   1.767  0.07764 .
## genreComedy                      0.98293    2.98616   0.329  0.74214
## genreDocumentary                16.62940    7.09879   2.343  0.01946 *
## genreDrama                      10.82022    2.53592   4.267 2.29e-05 ***
## genreHorror                     -6.85914    4.44890  -1.542  0.12364
## genreMusical & Performing Arts  19.77580    6.05167   3.268  0.00114 **
## genreMystery & Suspense          1.07470    3.33231   0.323  0.74717
## genreOther                      11.97678    5.04504   2.374  0.01790 *
## genreScience Fiction & Fantasy  -4.02326    6.35636  -0.633  0.52700
## runtime                          0.17768    0.04245   4.185 3.26e-05 ***
## mpaa_ratingNC-17               -10.41867   13.53800  -0.770  0.44183
## mpaa_ratingPG                   -9.76940    4.90043  -1.994  0.04663 *
## mpaa_ratingPG-13               -15.63729    5.00355  -3.125  0.00186 **
## mpaa_ratingR                    -9.99881    4.85316  -2.060  0.03979 *
## mpaa_ratingUnrated              -6.48384    5.60508  -1.157  0.24780
## thtr_rel_month                  -0.07782    0.20458  -0.380  0.70379
## best_actor_winyes               -1.00440    2.11506  -0.475  0.63504
## best_actress_winyes             -0.11666    2.32657  -0.050  0.96002
## best_dir_winyes                  6.15676    2.92599   2.104  0.03576 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.83 on 627 degrees of freedom
## Multiple R-squared:  0.2495, Adjusted R-squared:  0.2231
## F-statistic: 9.474 on 22 and 627 DF,  p-value: < 2.2e-16
```

We will use a backward elimination p-value method to arrive at our parsimonious model. I chose this method because, with so many predictors under consideration, the adjusted R squared methods would be unwieldy.

It would appear that whether or not a movie features a woman who has won best actress has the greatest p-value, so we will eliminate that variable first.

```
m1<-lm(audience_score~title_type+genre+runtime+mpaa_rating+thtr_rel_month+best_actor_win+best_dir_win, d
summary(m1)
```

```
##
## Call:
## lm(formula = audience_score ~ title_type + genre + runtime +
##     mpaa_rating + thtr_rel_month + best_actor_win + best_dir_win,
##     data = movies)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -52.66 -12.99   1.23  13.26  39.50
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     58.08741    8.97817   6.470 1.98e-10 ***
## title_typeFeature Film         -11.83355    6.60648  -1.791  0.07374 .
## title_typeTV Movie             -21.41802   10.40349  -2.059  0.03993 *
## genreAnimation                   4.65993    6.95709   0.670  0.50323
## genreArt House & International   9.51736    5.38166   1.768  0.07747 .
## genreComedy                      0.96844    2.96978   0.326  0.74446
## genreDocumentary                16.62061    7.09099   2.344  0.01939 *
## genreDrama                      10.80326    2.51126   4.302 1.96e-05 ***
## genreHorror                     -6.86630    4.44308  -1.545  0.12275
## genreMusical & Performing Arts  19.77684    6.04683   3.271  0.00113 **
## genreMystery & Suspense          1.05620    3.30919   0.319  0.74970
## genreOther                      11.96525    5.03579   2.376  0.01780 *
## genreScience Fiction & Fantasy  -4.02515    6.35120  -0.634  0.52647
## runtime                          0.17738    0.04199   4.224 2.76e-05 ***
## mpaa_ratingNC-17               -10.39950   13.52185  -0.769  0.44213
## mpaa_ratingPG                   -9.77016    4.89651  -1.995  0.04644 *
## mpaa_ratingPG-13               -15.63761    4.99957  -3.128  0.00184 **
## mpaa_ratingR                    -9.99467    4.84861  -2.061  0.03968 *
## mpaa_ratingUnrated              -6.47792    5.59938  -1.157  0.24775
## thtr_rel_month                  -0.07793    0.20441  -0.381  0.70316
## best_actor_winyes               -1.01135    2.10883  -0.480  0.63169
## best_dir_winyes                  6.15251    2.92244   2.105  0.03566 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.82 on 628 degrees of freedom
## Multiple R-squared:  0.2495, Adjusted R-squared:  0.2244
## F-statistic: 9.941 on 21 and 628 DF,  p-value: < 2.2e-16
```

Next we will eliminate theatrical release month. (Even though genre Comedy has a higher p-value, some of the genres are still statistically significant, so we cannot eliminate that predictor.)

```
m2<-lm(audience_score~title_type+genre+runtime+mpaa_rating+best_actor_win+best_dir_win, data=movies)
summary(m2)
```

```
##
## Call:
```

```
## lm(formula = audience_score ~ title_type + genre + runtime +
##     mpaa_rating + best_actor_win + best_dir_win, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.902 -13.126   1.301  13.301  39.384
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    57.90767    8.95968   6.463 2.06e-10 ***
## title_typeFeature Film        -11.81662    6.60184  -1.790  0.07395 .
## title_typeTV Movie            -21.29456   10.39138  -2.049  0.04085 *
## genreAnimation                  4.56065    6.94749   0.656  0.51178
## genreArt House & International  9.51314    5.37800   1.769  0.07739 .
## genreComedy                     0.92043    2.96510   0.310  0.75634
## genreDocumentary               16.62488    7.08616   2.346  0.01928 *
## genreDrama                     10.81328    2.50942   4.309 1.90e-05 ***
## genreHorror                    -6.87936    4.43992  -1.549  0.12178
## genreMusical & Performing Arts 19.75002    6.04231   3.269  0.00114 **
## genreMystery & Suspense         1.12324    3.30227   0.340  0.73386
## genreOther                     12.04755    5.02774   2.396  0.01686 *
## genreScience Fiction & Fantasy -4.00936    6.34675  -0.632  0.52780
## runtime                         0.17394    0.04099   4.244 2.53e-05 ***
## mpaa_ratingNC-17              -10.27603   13.50878  -0.761  0.44713
## mpaa_ratingPG                  -9.78875    4.89294  -2.001  0.04587 *
## mpaa_ratingPG-13              -15.58932    4.99457  -3.121  0.00188 **
## mpaa_ratingR                  -10.00672    4.84521  -2.065  0.03931 *
## mpaa_ratingUnrated             -6.44191    5.59478  -1.151  0.25000
## best_actor_winyes              -1.04551    2.10550  -0.497  0.61967
## best_dir_winyes                 6.13159    2.91994   2.100  0.03613 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.81 on 629 degrees of freedom
## Multiple R-squared:  0.2493, Adjusted R-squared:  0.2254
## F-statistic: 10.44 on 20 and 629 DF,  p-value: < 2.2e-16
```

Best actor winner has the highest of the remaining p-values, so we'll eliminate that one.

```
m3<-lm(audience_score~title_type+genre+runtime+mpaa_rating+best_dir_win, data=movies)
summary(m3)
```

```
##
## Call:
## lm(formula = audience_score ~ title_type + genre + runtime +
##     mpaa_rating + best_dir_win, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.671 -13.017   1.343  13.316  39.399
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                     58.3175     8.9163   6.541 1.27e-10 ***
## title_typeFeature Film         -11.8562     6.5974  -1.797  0.07280 .
## title_typeTV Movie            -21.2172    10.3840  -2.043  0.04144 *
## genreAnimation                  4.4799     6.9414   0.645  0.51891
## genreArt House & International  9.5847     5.3728   1.784  0.07492 .
## genreComedy                     0.9012     2.9631   0.304  0.76111
## genreDocumentary               16.5874     7.0815   2.342  0.01947 *
## genreDrama                     10.7557     2.5052   4.293 2.04e-05 ***
## genreHorror                    -6.8424     4.4366  -1.542  0.12352
## genreMusical & Performing Arts 19.7933     6.0381   3.278  0.00110 **
## genreMystery & Suspense         0.9625     3.2844   0.293  0.76958
## genreOther                     11.9933     5.0236   2.387  0.01726 *
## genreScience Fiction & Fantasy -3.9078     6.3397  -0.616  0.53785
## runtime                         0.1696     0.0400   4.239 2.58e-05 ***
## mpaa_ratingNC-17              -10.6730    13.4770  -0.792  0.42869
## mpaa_ratingPG                  -9.8586     4.8880  -2.017  0.04413 *
## mpaa_ratingPG-13              -15.6005     4.9915  -3.125  0.00186 **
## mpaa_ratingR                  -10.0048     4.8423  -2.066  0.03923 *
## mpaa_ratingUnrated             -6.4048     5.5909  -1.146  0.25241
## best_dir_winyes                 6.0833     2.9166   2.086  0.03740 *
## ---
## 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:  0.249,  Adjusted R-squared:  0.2264
## F-statistic: 10.99 on 19 and 630 DF,  p-value: < 2.2e-16
```

At this point, at least one level of all of our remaining variables (title type, genre, runtime, mpaa rating and best director win) is statistically significant, so this will be our final model.

All other variables being equal we find that:

1. Having a director who has won an academy award yields an audience rating that is 6 points higher on average than a non-academy award winning director.
2. Documentaries tend to score 21 points higher than TV movies and 11.8 points higher than feature films.
3. The genre with the highest audience score is Musical & Performing Arts, which sees a boost of 19.8 points above baseline. The genre with the most negative impact on average audience score is horror, which is 6.8 points below baseline.
4. Longer movies tend to get higher audience ratings than shorter movies, with every additional minute increasing the predicted audience score by 0.17 on average.
5. G-rated movies tend to get the highest audience ratings, with unrated movies trailing an average of 6 points behind, followed by PG, R and NC-17 movies, which all tend to be rated around 10 points lower than G movies. The lowest ratings are given to PG-13 movies, about 15.6 points lower than G.

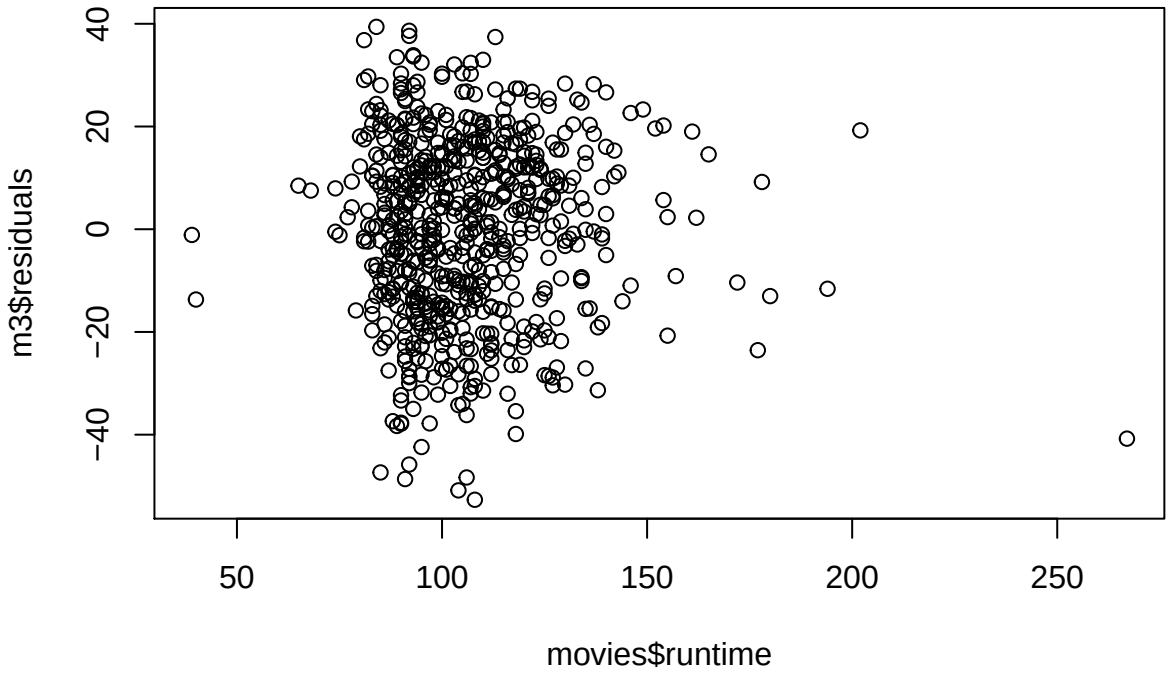
Adjusted R-squared = 0.2264, so 22.64% of the variability in audience ratings are accounted for by our regression model.

###Diagnostics for MLR

Now let's check model diagnostics for the predictors that remain.
First, we must check that each numerical predictor has a linear relationship with y. The only remaining

numerical predictor is runtime. We saw in our EDA that the scatterplot of audience score vs. runtime has a low correlation and no obvious departures from linearity. Let's also look at a residual plot of residuals vs. runtime in order to take other predictors into account.

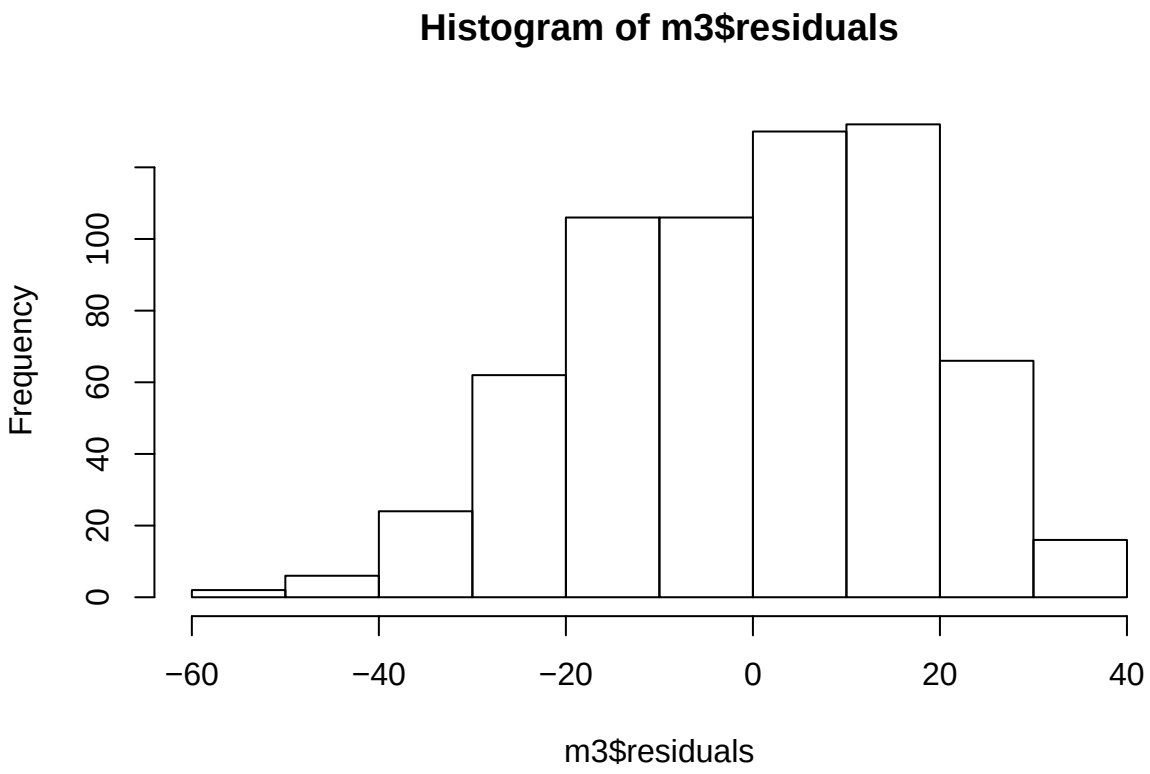
```
plot(m3$residuals~movies$runtime)
```



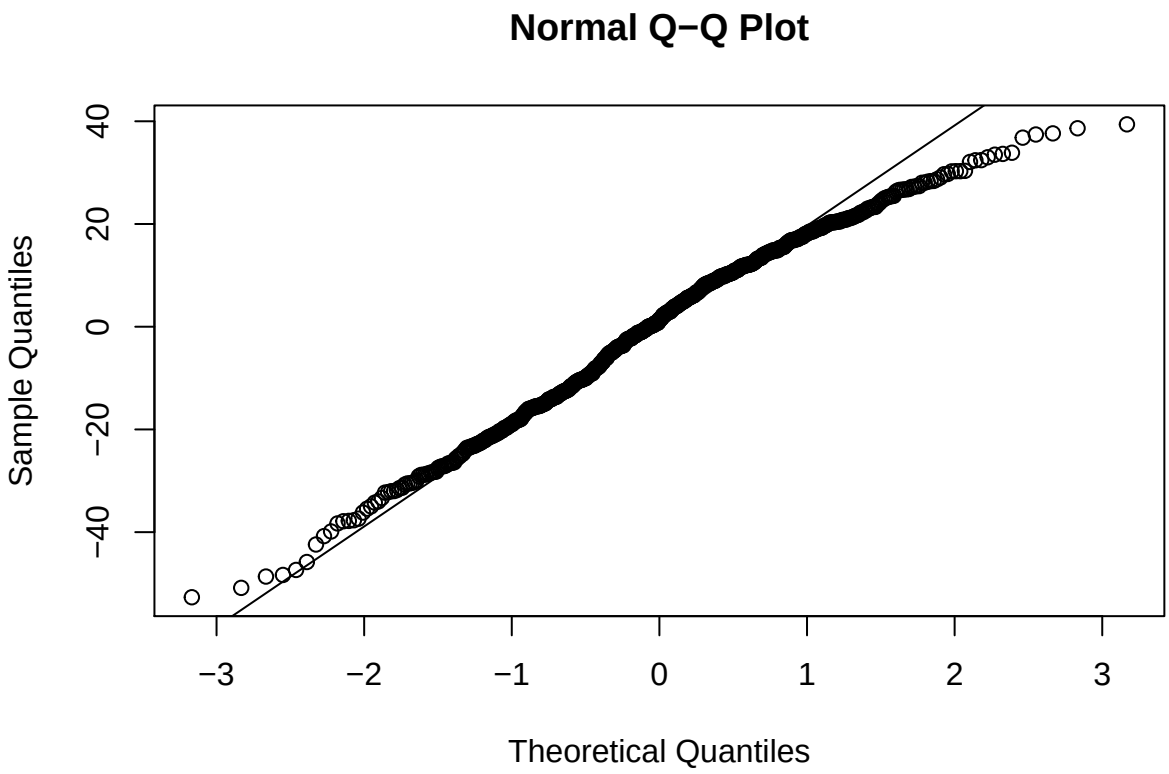
The residual plot shows no clear form, so we should be all set for this condition.

Next, we want to check the normality of our residuals.

```
hist(m3$residuals)
```



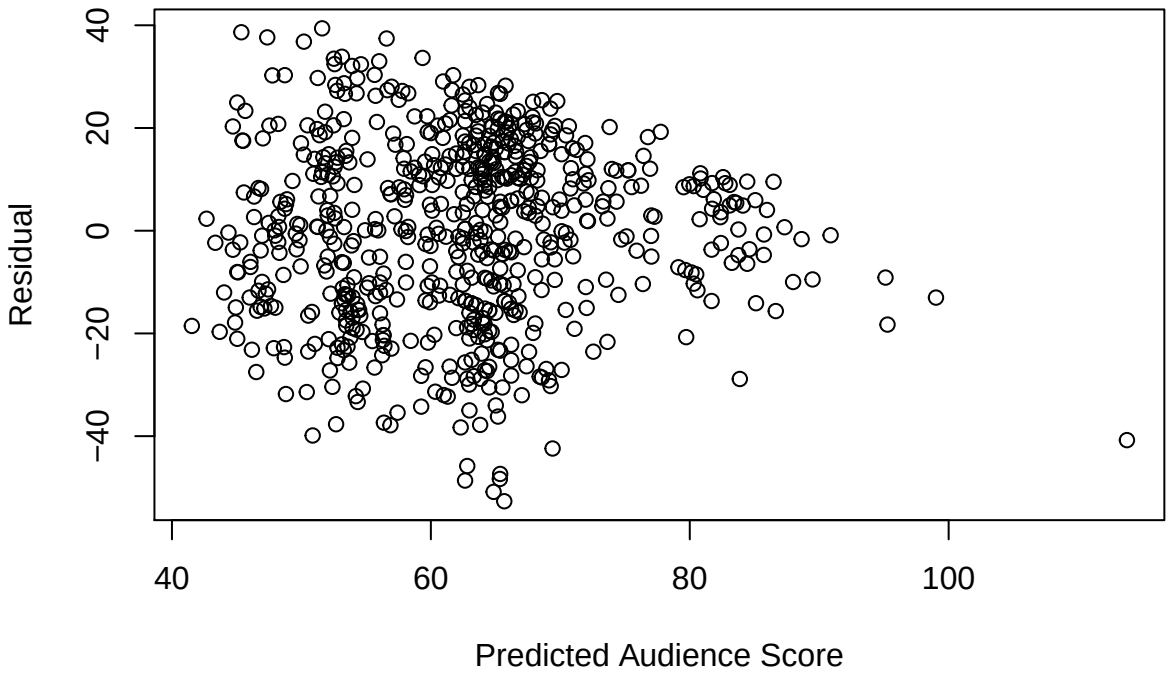

```
qqnorm(m3$residuals)
qqline(m3$residuals)
```



The histogram and Normal Quantile Plot of the residuals both show a slight left skew in the residuals, but neither shows an alarming departure from Normality.

Next we want to check the constant variability of the residuals. We will check this by making a plot of residuals vs. predicted audience score, so that all predictors are considered.

```
plot(m3$residuals~m3$fitted, xlab="Predicted Audience Score", ylab="Residual")
```



We have approximately equal variability in residuals when predicted audience scores are between 45 and 75. There were far fewer cases for which the predicted score was above 75, so it is hard to say whether the variability for those predictions would be equal to that of the lower predictions.

Part 5: Prediction

I have selected the movie Whiskey Tango Foxtrot (2016) to use for my prediction. The actual audience score for the movie is 55 according to https://www.rottentomatoes.com/m/whiskey_tango_foxtrot. Also on Rotten Tomatoes we find that it is a feature film, genre is comedy, rating is R and Runtime is 111 minutes. It has two directors, Glenn Ficarra and John Requa, neither of whom has won an Oscar for best director, according to IMDB. http://www.imdb.com/name/nm0720135/awards?ref_=m_nm_awd&mode=desktop http://www.imdb.com/name/nm0275629/awards?ref_=nm_ql_2

```
WTF<-data.frame(genre="Comedy", mpaa_rating="R", runtime=111, best_dir_win="no", title_type="Feature Fi
predict(m3, WTF)
```

```
##        1
## 56.17901
```

Predicted audience score differed from actual audience score by just 1.179, which is surprisingly close, considering that the standard error of the residuals is 17.8.

```r
predict(m3, WTF, interval = "prediction", level=0.95)
```

```
##        fit      lwr      upr
## 1 56.17901 20.95208 91.40594
```

We are, therefore, 95% confident that the true audience score for this movie will be between 20.952 and 91.406. The large standard error of the residuals results in a very wide margin, which easily captures the true value of audience score.

Part 6: Conclusion

Bearing in mind that the following traits cannot be said to cause a higher audience rating, our analysis suggests that the following qualities are associated with higher audience ratings: documentary, musical, rated G, Oscar-winning director, longer runtime.

It is important to note that better audience scores of a movie do not guarantee that a movie will be more financially successful. It would be useful to have information about the amount that these movies cost and the amount that they made in the box office and through follow-up sales and licensing to provide better advice about the types of projects you should pursue in the future to maximize financial success.