Stat 706 Final Project: Investigating Predictors of Movie Profitability

Nickhil Sethi

TODO

- Run plots of relevant variables against revenue. covariance matrix of genres.
- Run regression using step function AIC information criterion. Identify meaningful values using p-values.
- Discuss the results. Run diagnostics, were the assumptions of LR correct?
- Prediction intervals

Introduction

The goal of this paper is to investigate film profitability in Kaggle's "The Movies Dataset", an aggregate dataset from GroupLens and TMDB containing information on revenue, budget, ratings, and various qualitative traits (e.g. genre, medium) for roughly ~30,000 films.

In particular, we hope to answer the following questions:

- Which genres are most profitable?
- How is critical reception related to profitability?
- How is release date related to profitability?
- How is budget related to profitability?

We begin from a complete set of predictors and use an iterative procedure to minimize the Akaike-Information Criterion (i.e. R's `step` function) in order to discover the most powerful predictors. We then examine the reduced model for significance, interpret the results, and construct prediction intervals for movies of varying characteristics.

Data

Transformations, Cleaning, and Schema

The Movies dataset contains three tables relevant to our analysis – `movies_metadata`, `ratings`, and a join table `links`. The three tables have the following schemas:

```
MOVIES_METADATA {
  genres: [{ genreId: <int>,  name: <str> }],
  revenue: float,
  budget: float,
  imdbId: int
}

RATINGS {
  movieId: int,
  userId: int,
  rating: float,
}
```

```
LINKS {
  movieId: int,
  imdbId: int
}
```

These three tables were transformed into a single table `movies` used in the analysis for this paper. The schema of `movies` is as follows:

```
movies {
  imdb_id: int,
  budget: float,
  profit: float,
  release_date: date,
  average_rating: float,
  genre_action: boolean,
  genre_thriller: boolean,
  ...
  genre_music: boolean
}
```

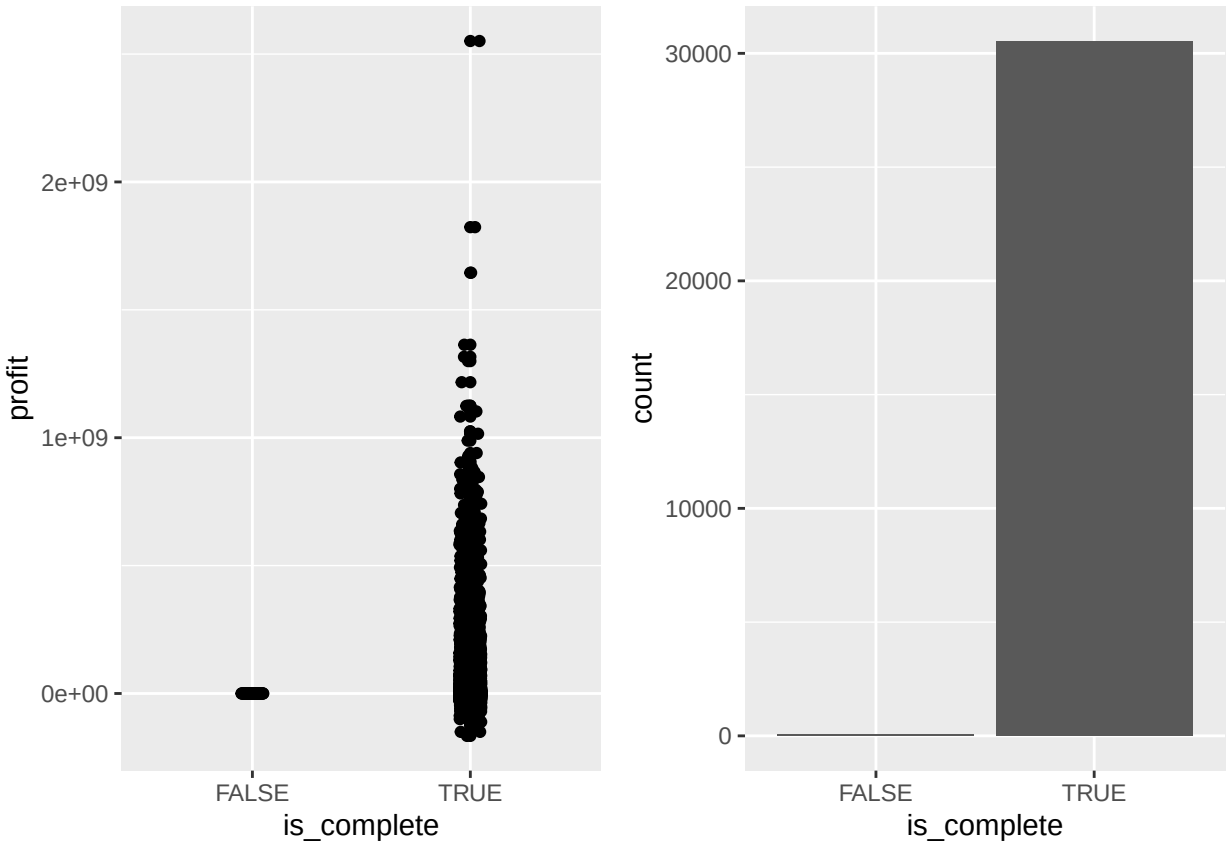
Several transformations were performed to turn the three original tables into the usable format contained in `movies`.

1. The genre column of `movies_metadata` is JSON, encoded as a list of pairs e.g. `[{genreId: 1, genreName: Action}, {genreId: 2, genreName: Comedy}]`, with each pair representing a genre the film is associated with; this column was converted to a set of boolean columns (e.g. `genre_action`) with `True` representing that the film belongs to that genre.
2. The `ratings` table contains movie ratings at the level of `(movieId, userId)` pairs, i.e. at the level of an individual critic's review; a grouping operation was performed to compute the average rating for each movie.
3. A join was performed between `movies_metadata` and the grouped version of `ratings` using the join table `links`.
4. The `profit` column was added, defined as `revenue - budget` for each row; note the assumption here that all films use exactly there budget.
5. Finally, rows with missing values are dropped. As can be seen from the chart below, the variable `is_complete` may be correlated with `profit`, (with incomplete rows typically having `0` profit); however, only a tiny fraction of rows are missing data, and this is not likely to influence the results in any major way.

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```

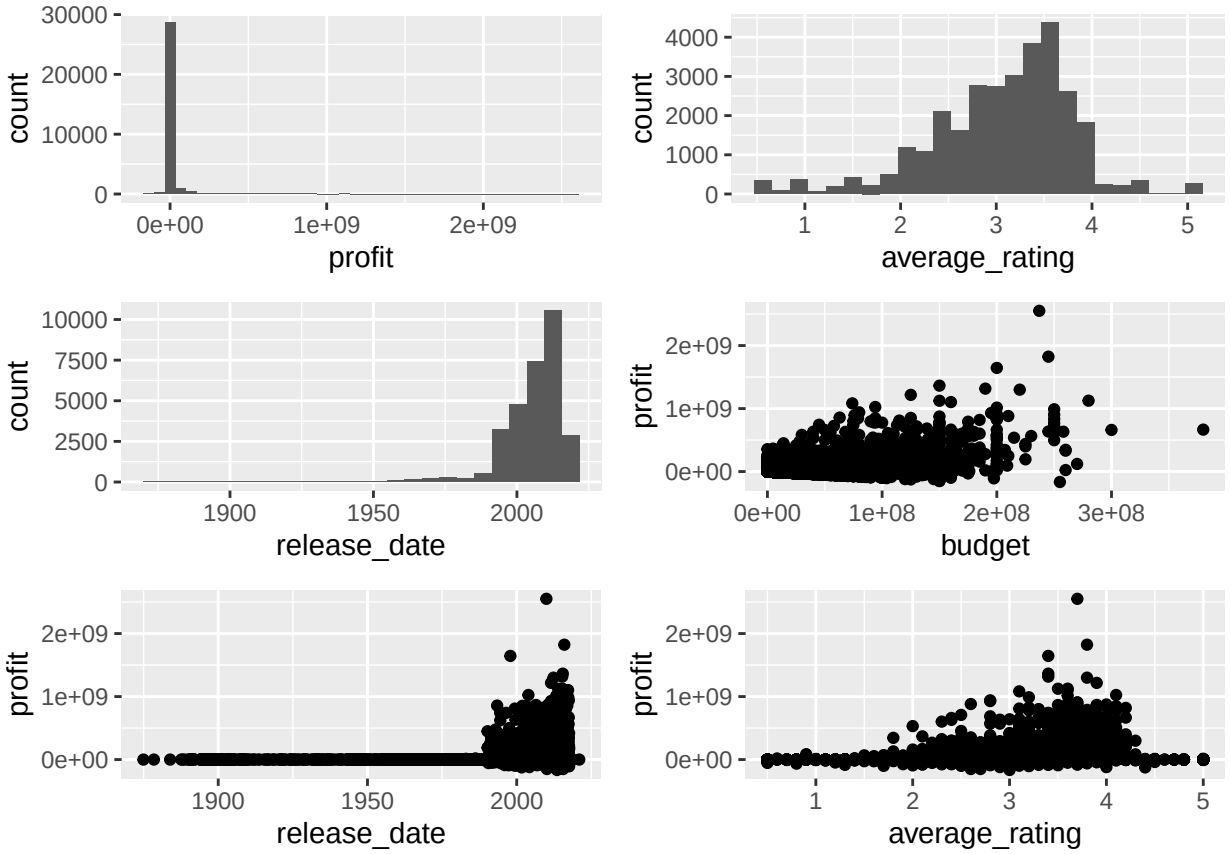
```
## Warning: Removed 3 rows containing missing values (geom_point).
```



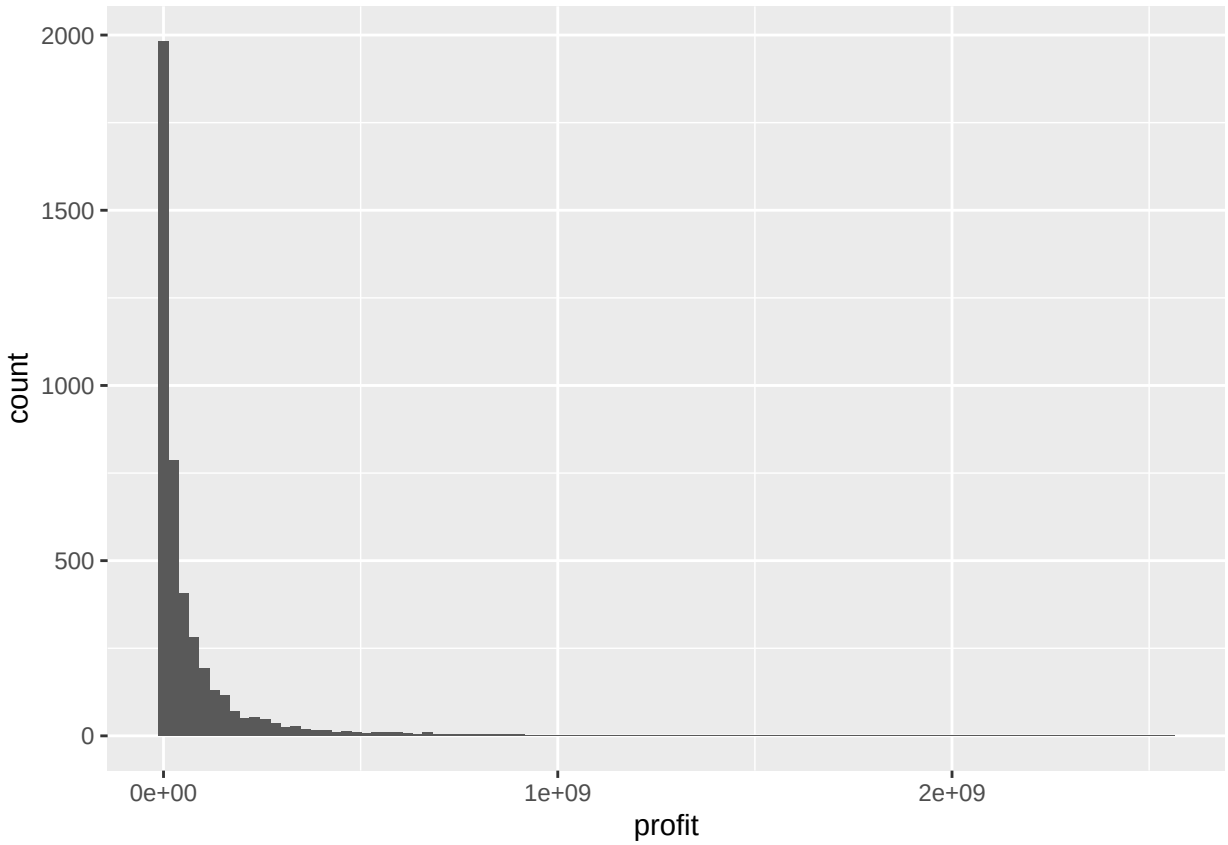
Description of Data and Considerations for Analysis

The `movies` table contains four continuous variables – `profit`, `budget`, `average_rating`, and `release_date`, as well as 21 boolean variables reperesenting genre classifications e.g. `genre_action`.

The distributions of three continuous variables `release_date`, `average_rating`, and `budget`, as well as scatter plots of these three variables against `profit` are shown below:



We can glean a few interesting facts from inspection of the above graphs. Most of these variables appear to be normally distributed, with the potential exception of `profit` which is (unfortunately) heavily peaked at zero. The non-normality of `profit` becomes more visible if we restrict to films that are profit-positive, where we see the distribution more closely resembles an exponential than a gaussian:



Since our predictors are normally distributed but our response variable is not, our errors are likely not normally distributed as well (this is actually confirmed in the `Discussion` section). In the case of non-normally distributed errors, p-values and confidence intervals may not be accurate and are difficult to interpret.

A second issue results from the fact that `budget`, `average_rating`, and `release_date` are all correlated with profit, and are therefore correlated with each other; our procedure for model selection is minimizing AIC, but this may not yield the most interpretable model.

Methods and Results

We begin by fitting a model with all available predictors; `budget`, `average_rating`, `release_date`, and every genre are included in the model. Then, we call `R`'s `step` function on the model, which iteratively drops variables by testing if removing them from the model lowers the `AIC`.

Below is a summary of the results for the complete model:

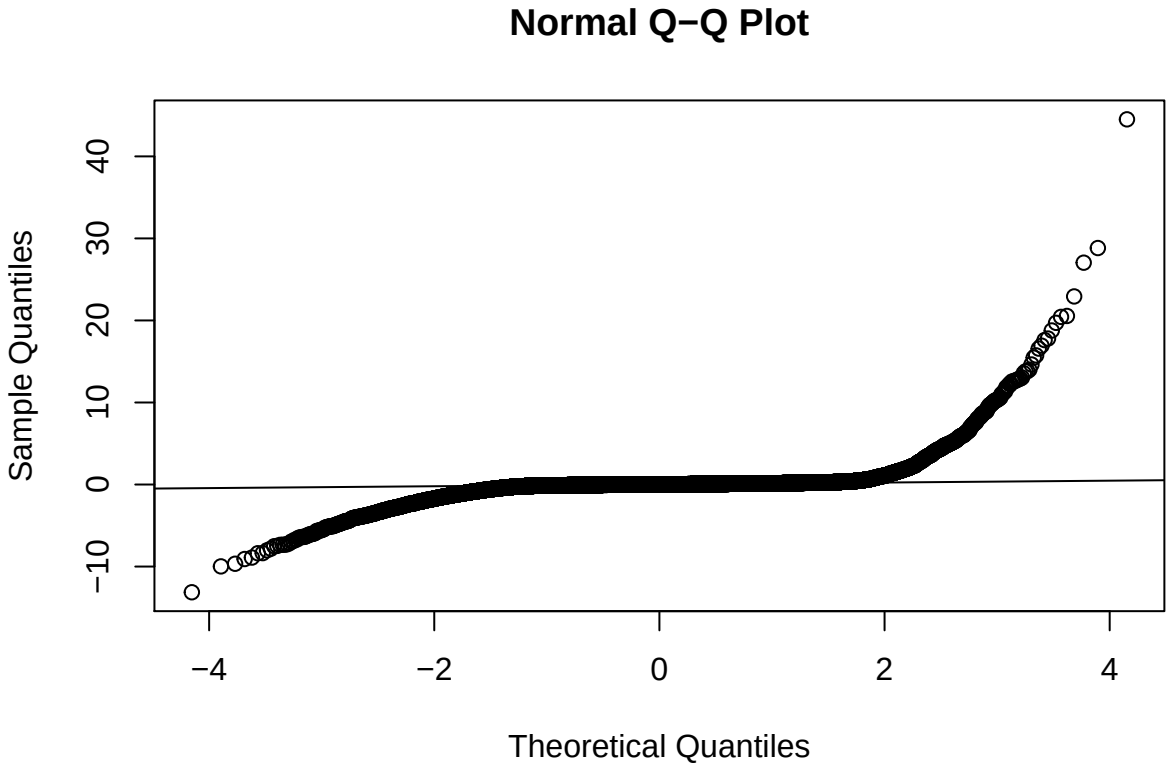
```
##
## Coefficients: (2 not defined because of singularities)
##                              Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)               -1.3446e+07  1.5025e+06  -8.9488 < 2.2e-16
## budget                     1.8426e+00  1.4143e-02 130.2893 < 2.2e-16
## release_date               9.6058e+01  6.0806e+01   1.5797 0.1141787
## average_rating             4.1698e+06  3.8602e+05  10.8021 < 2.2e-16
## genre_thrillerTRUE        -3.4121e+06  8.2653e+05  -4.1282 3.666e-05
## genre_warTRUE             -7.1465e+06  2.0610e+06  -3.4676 0.0005259
## genre_fantasyTRUE          1.8692e+06  1.2951e+06   1.4433 0.1489442
## genre_documentaryTRUE     -1.4743e+06  1.0178e+06  -1.4485 0.1474813
## genre_actionTRUE          -3.6248e+06  8.8509e+05  -4.0953 4.226e-05
## genre_familyTRUE           1.6498e+06  1.2082e+06   1.3655 0.1721049
```

```
## genre_crimeTRUE           -2.3810e+06  1.0542e+06  -2.2586 0.0239150
## genre_comedyTRUE          -1.1520e+06  6.7057e+05  -1.7180 0.0858156
## genre_foreignTRUE          2.2608e+06  1.3653e+06   1.6560 0.0977408
## genre_adventureTRUE        5.3883e+06  1.2058e+06   4.4689 7.892e-06
## genre_mysteryTRUE         -1.5966e+06  1.3192e+06  -1.2103 0.2261867
## genre_science_fictionTRUE -1.1232e+06  1.1516e+06  -0.9754 0.3293717
## genre_horrorTRUE           2.9958e+06  1.0137e+06   2.9554 0.0031251
## genre_dramaTRUE           -2.9028e+06  6.4744e+05  -4.4835 7.369e-06
## genre_animationTRUE        4.0421e+06  1.3589e+06   2.9745 0.0029366
## genre_romanceTRUE          1.1204e+06  8.3058e+05   1.3489 0.1773808
## genre_westernTRUE         -1.6938e+07  3.2277e+06  -5.2477 1.550e-07
## genre_historyTRUE         -8.0749e+06  1.7238e+06  -4.6844 2.819e-06
## genre_tv_movieTRUE         7.6480e+05  1.9200e+06   0.3983 0.6903861
## genre_musicTRUE            2.0358e+05  1.6379e+06   0.1243 0.9010805
##
## n = 30542, p = 24, Residual SE = 47468012.77514, R-Squared = 0.4

## [1] AIC for full model: 1166394.743765
```

And see here for the summary results of the reduced model:

```
##                         Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)         -1.3909e+07  1.4602e+06  -9.5253 < 2.2e-16
## budget               1.8421e+00  1.4041e-02 131.1933 < 2.2e-16
## release_date         8.6901e+01  6.0599e+01   1.4340 0.1515726
## average_rating       4.0986e+06  3.7942e+05  10.8024 < 2.2e-16
## genre_thrillerTRUE  -3.3626e+06  7.8356e+05  -4.2915 1.781e-05
## genre_warTRUE       -6.9866e+06  2.0579e+06  -3.3950 0.0006871
## genre_fantasyTRUE    1.8666e+06  1.2875e+06   1.4498 0.1471351
## genre_actionTRUE    -3.5369e+06  8.6294e+05  -4.0987 4.165e-05
## genre_familyTRUE     1.7142e+06  1.1965e+06   1.4327 0.1519546
## genre_crimeTRUE     -2.4284e+06  1.0430e+06  -2.3284 0.0198960
## genre_foreignTRUE    2.3254e+06  1.3628e+06   1.7064 0.0879511
## genre_adventureTRUE  5.3819e+06  1.2029e+06   4.4742 7.698e-06
## genre_horrorTRUE     3.0852e+06  9.8701e+05   3.1258 0.0017751
## genre_dramaTRUE     -2.2741e+06  5.7846e+05  -3.9313 8.469e-05
## genre_animationTRUE  4.2049e+06  1.3414e+06   3.1347 0.0017220
## genre_westernTRUE   -1.6616e+07  3.2228e+06  -5.1557 2.543e-07
## genre_historyTRUE   -7.9921e+06  1.7160e+06  -4.6574 3.215e-06
##
## n = 30542, p = 17, Residual SE = 47468590.08865, R-Squared = 0.4

## [1] AIC for reduced model: 1166388.491376
```

The AIC minimzation process results in seven genre variables being dropped, yielding a decrease in AIC from 1166395 to 1166388. Both models have an R^2 of .4.



Discussion

Model formulation

The model formulation and technique used here was chosen to "let the data speak"; that is, to determine the best predictors of `profit` without imposing too much speculation on model structure from the outset. For this reason, interaction effects between genres e.g. "action-comedy" or "musical-drama" were not accounted for. Doing so would entail arbitrary judgements about which interactions are valid "compound" genres; on the other hand, adding all pairwise interactions among 20 genres would add 190 additional predictors, with which comes numerous co-linearity problems and computational overhead.

Interpretation

Considerable information was gleaned from the data. The minimization procedure eliminates 7 genre variables from the complete model, with more niche genres such as `tv_movie` being removed for more mainstream ones such as `action`. Inspecting the output of the linear models above, we see the the predictors dropped by the `step` function typically had p-values well above .05 in the original model.

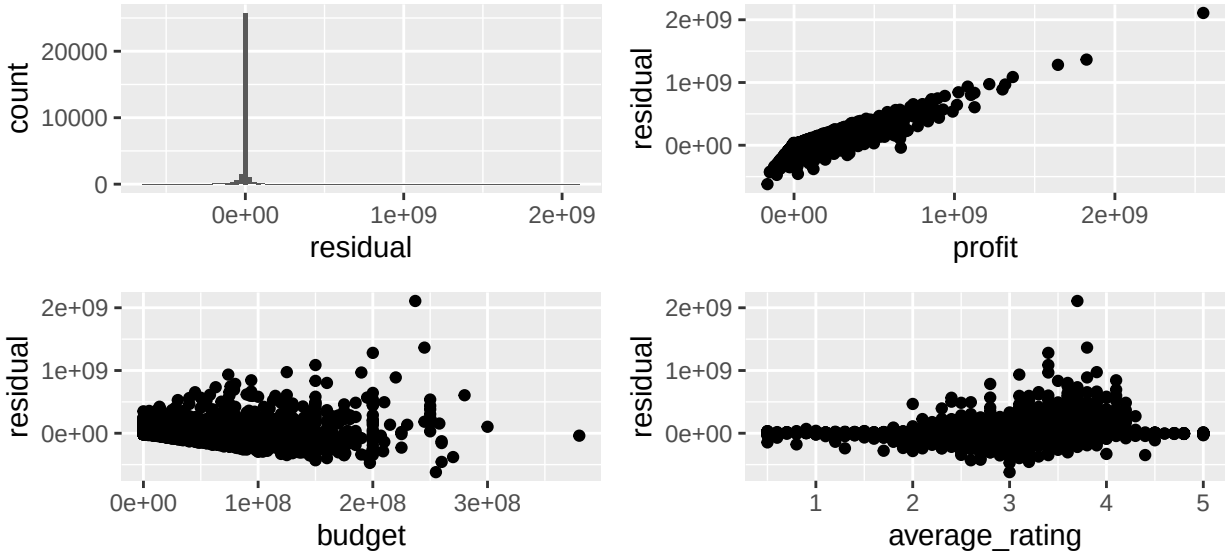
Genres `drama`, `war`, and `thriller` have negative effects on profit, indicating people avoid films with a negative emotional subtext. `history` and `western` have negative coefficients as well, perhaps indicating an audience preference for modern settings in film.

The `average_rating` variable has one of the strongest effects on `profit`, with a coefficient of `4.0986e+06`; a single star increase in rating entails an increase in profit of $4,000,000. Budget and release_date have moderate positive effects on profit, though the p-value on `release_date` of `.15` indicates a lack of statistical significance.

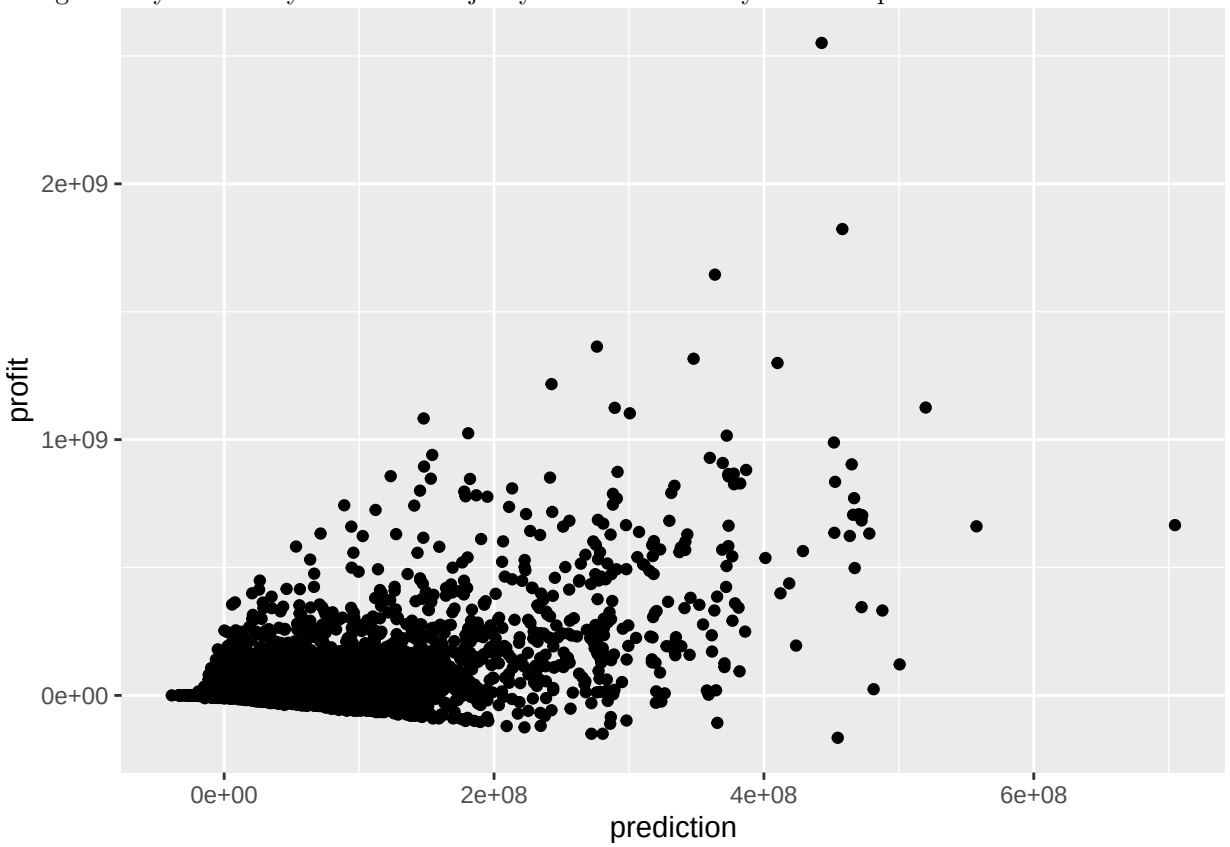
As alluded to earlier, the model errors are "heavy-tailed" rather than normally distributed. Though this has no pertinence to the sign and magnitude of our estimated coefficients, it does make our significance estimates and confidence intervals suspect. I attempted to solve this problem by restricting the analysis to films which

were profit positive and log-transforming the `profit` variable, but this did not remediate this issue.

As we can see below, the residual is very much correlated with all the continuous variables after the regression.

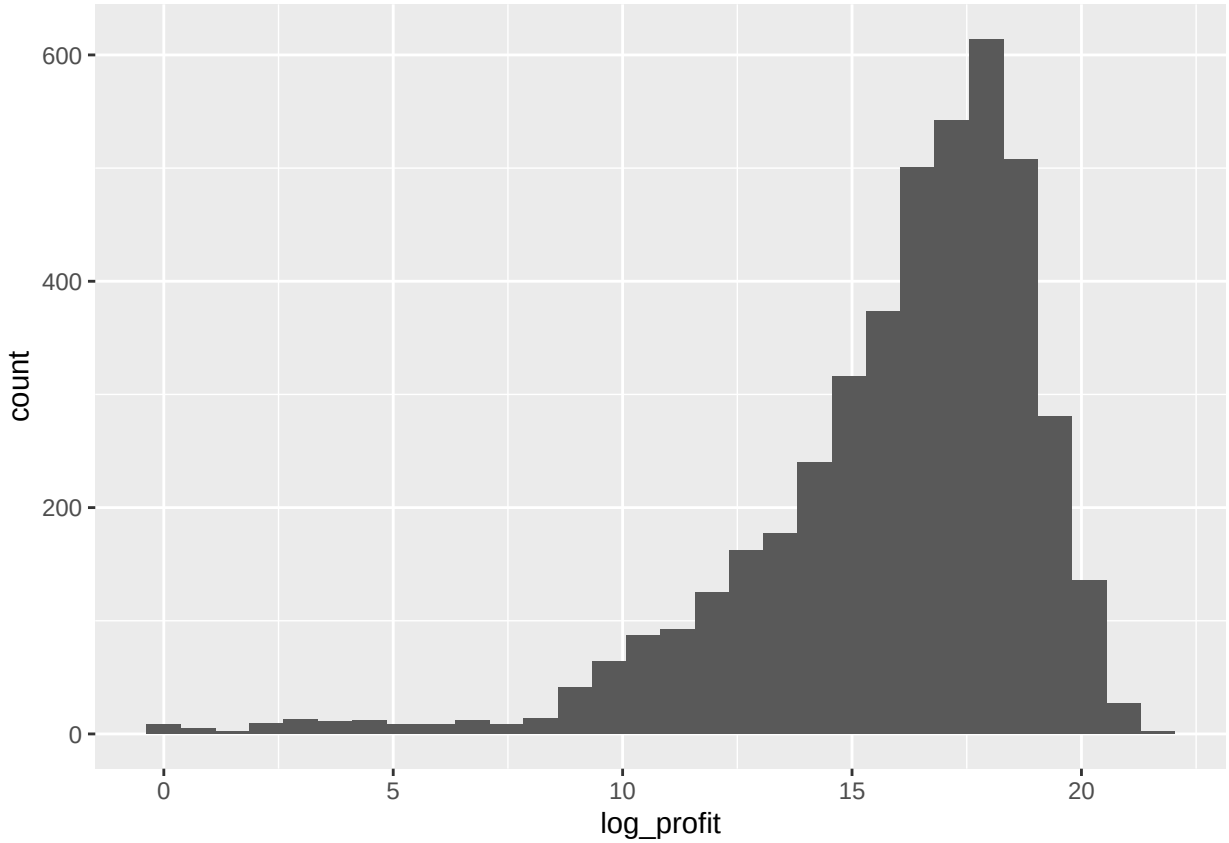


In fact the plots of above variables against the residuals closely resemble the plots of the variables against the response. This indicates the model is predicting zero for most films, likely being heavily driven by the vast majority of films which yield zero profit. As we can see below:



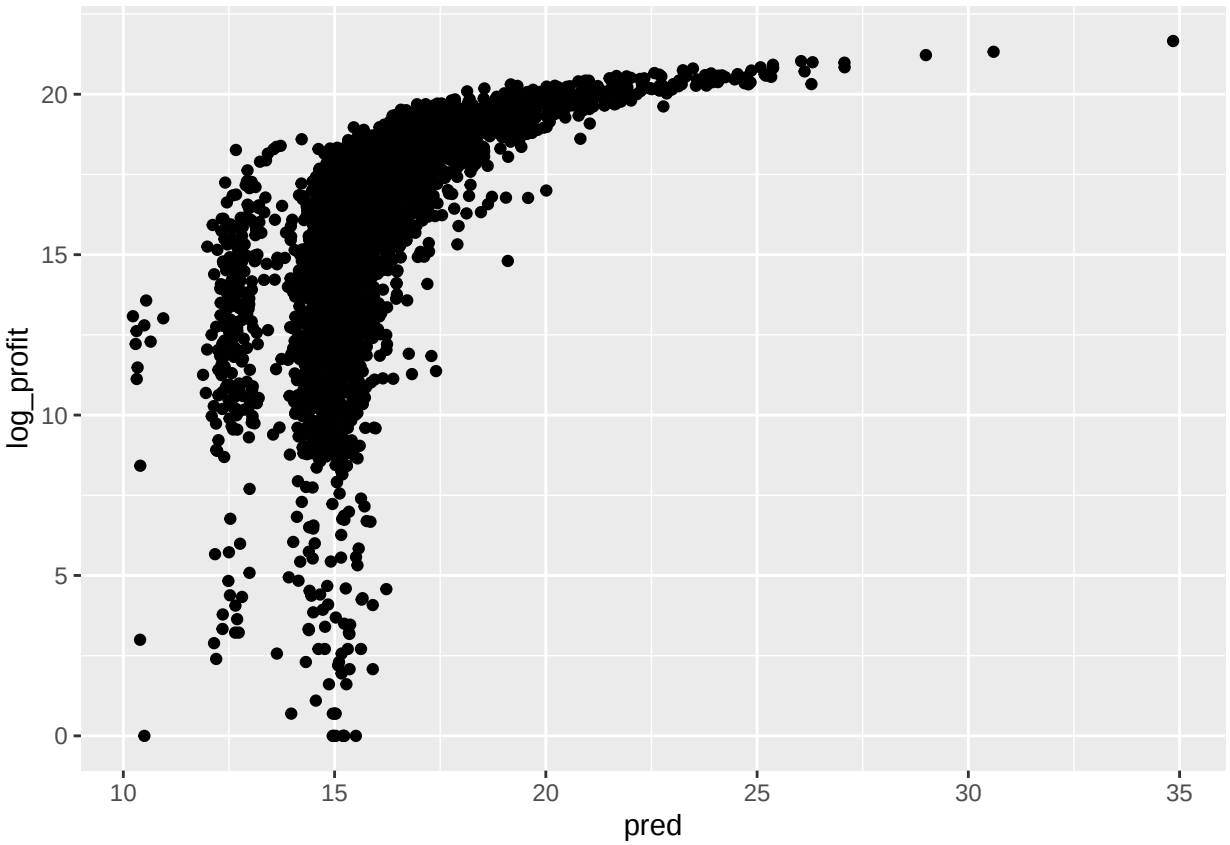
Let us briefly investigate a log transformation on the response. As we can see, dropping movies whose profit is not strictly greater than zero and then log transforming the profit column makes the data more closely resemble a gaussian:

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
##                           Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)            1.4094e+01  3.3146e-01  42.5215 < 2.2e-16
## profit                 5.9403e-09  3.6182e-10  16.4176 < 2.2e-16
## budget                 1.9992e-08  1.2550e-09  15.9307 < 2.2e-16
## release_date          -6.0640e-05  1.4219e-05  -4.2647 2.044e-05
## average_rating         4.7180e-01  8.2356e-02   5.7287 1.080e-08
## genre_thrillerTRUE     2.5214e-01  1.0681e-01   2.3607 0.0182839
## genre_warTRUE          4.3234e-01  2.5565e-01   1.6912 0.0908790
## genre_documentaryTRUE -2.3429e+00  2.0993e-01 -11.1604 < 2.2e-16
## genre_familyTRUE       6.3218e-01  1.3619e-01   4.6419 3.552e-06
## genre_crimeTRUE        4.4374e-01  1.2011e-01   3.6945 0.0002230
## genre_comedyTRUE       3.8936e-01  9.4583e-02   4.1166 3.916e-05
## genre_foreignTRUE     -2.1716e+00  3.2784e-01  -6.6241 3.916e-11
## genre_mysteryTRUE      4.3659e-01  1.5920e-01   2.7424 0.0061233
## genre_horrorTRUE       5.6954e-01  1.5037e-01   3.7875 0.0001542
## genre_dramaTRUE       -3.5836e-01  9.3840e-02  -3.8189 0.0001359
## genre_romanceTRUE      2.5088e-01  1.0531e-01   2.3823 0.0172487
## genre_historyTRUE      4.1948e-01  2.2256e-01   1.8848 0.0595249
##
## n = 4398, p = 17, Residual SE = 2.52213, R-Squared = 0.36
```

```
## [1] AIC for model with log transformation: 20637.175236
```



Future Directions and Improvements

Changing the link function to account for heavy tails.