Stat 706 Final Project: Investigating Predictors of Movie Profitability

Nickhil Sethi

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:

- How does genre affect profitability?
- 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 discuss future directions for this investigation.

Description of 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
}
```

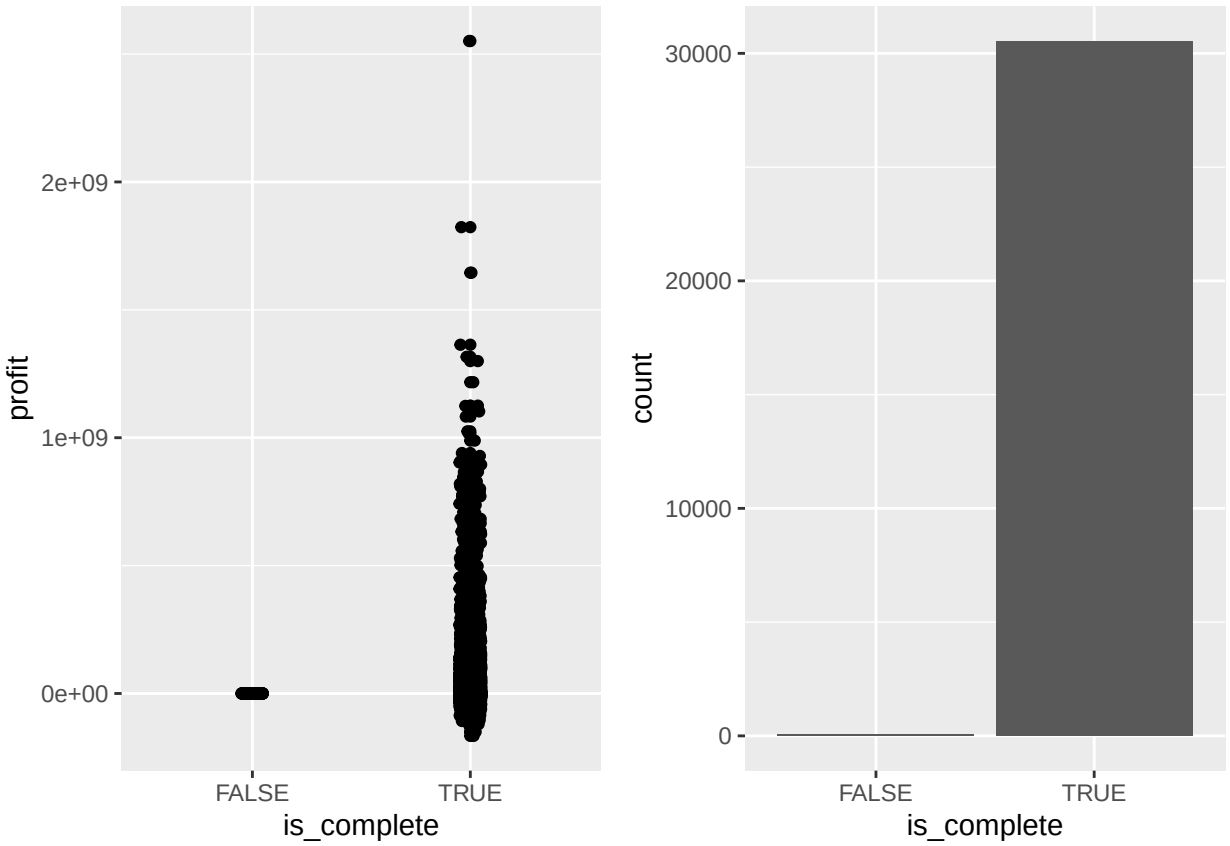
Several transformations were performed to turn the three original tables into the usable format contained in `movies`.

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
```

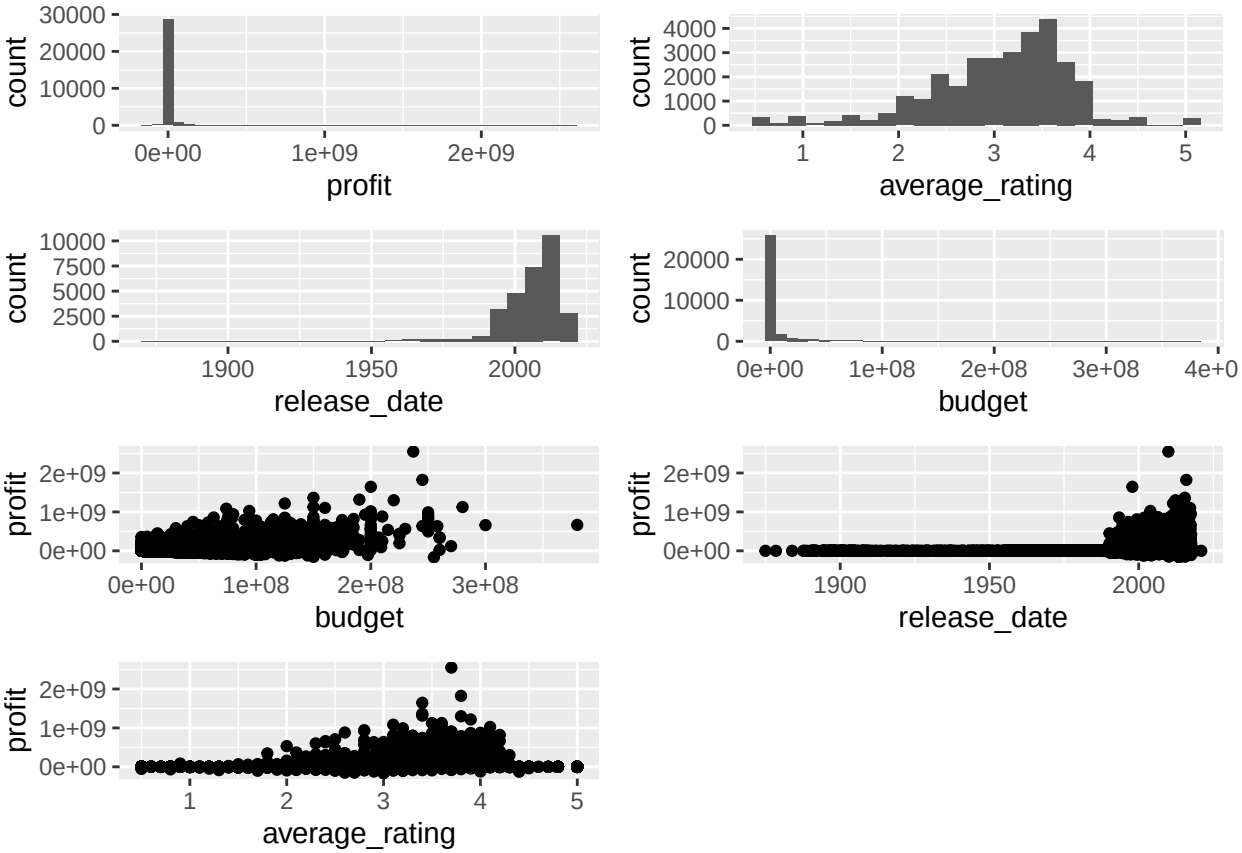
```
## Warning: Removed 3 rows containing missing values (geom_point).
```

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

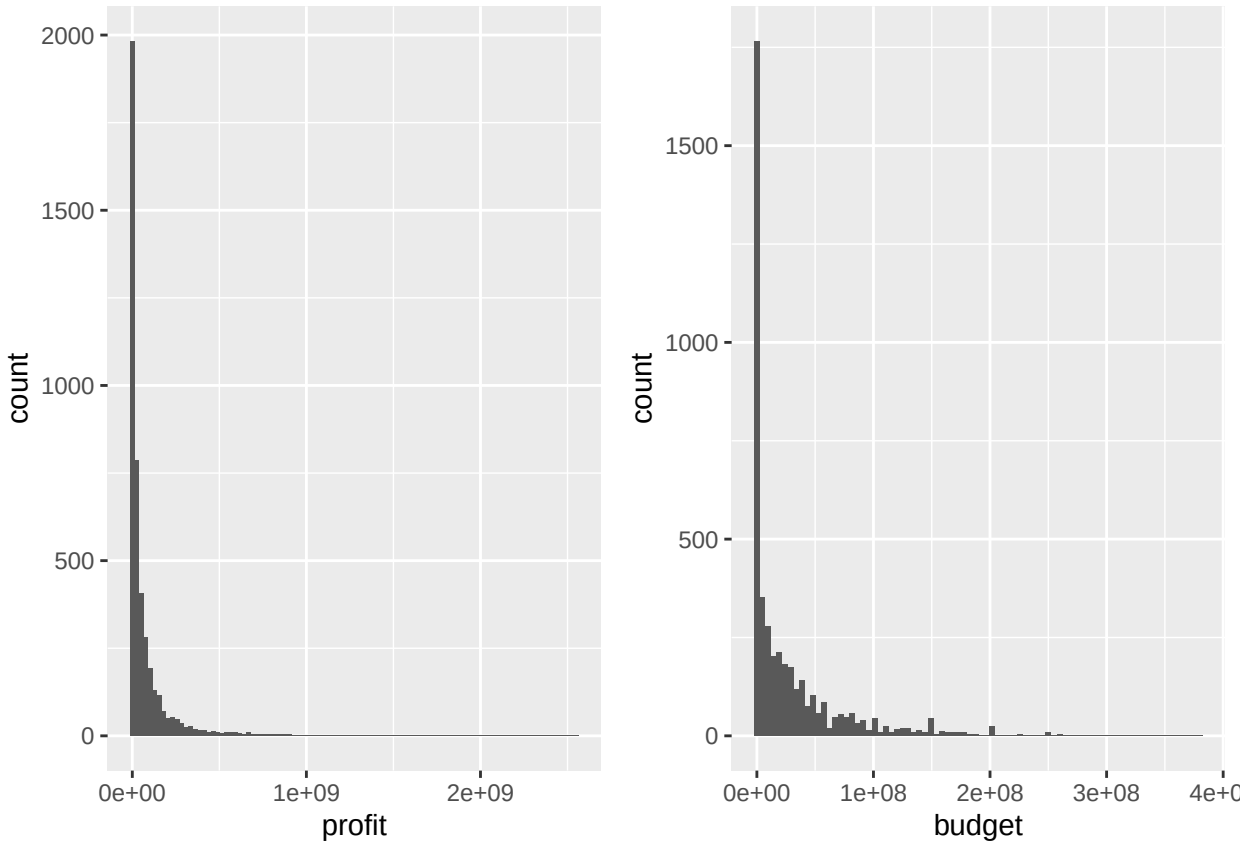


Distribution and Characteristics

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 the four continuous variables, as well as scatter plots of the input variables against `profit` are shown below:



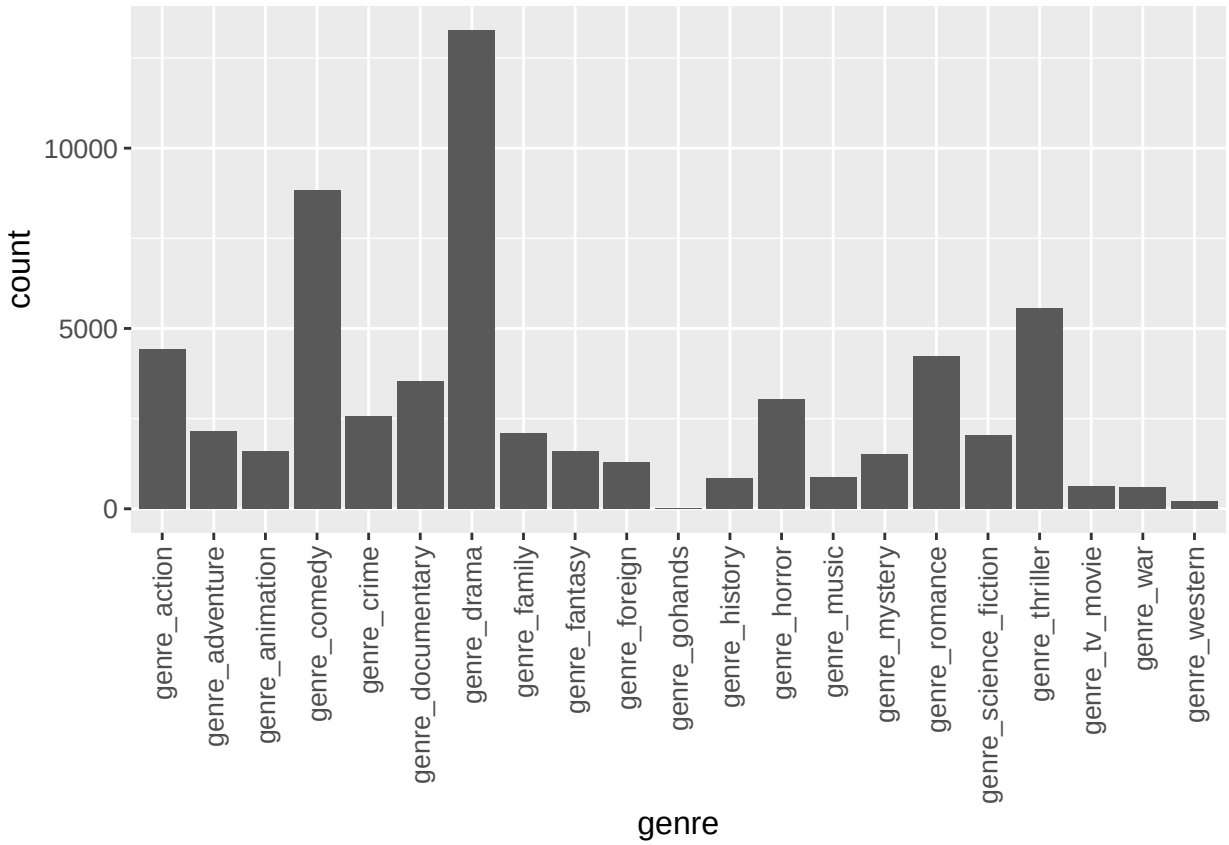
We can glean a few interesting facts from inspection of the above graphs. The continuous variables appear to be normally distributed, with the exceptions of `profit` and `budget` which are heavily peaked at zero. The non-normality of these two variables becomes more visible if we restrict to films that are profit-positive, where we see the distributions more closely resembles an exponential than a gaussian:



The heavy skew of these variables towards zero can pose a problem for linear regression. The response variable being peaked at zero encourages the model to simply output zero; furthermore, the input `budget` variable being peaked at zero reduces the certainty of the models predictions for high budget films. These concerns are actually confirmed in the `Discussion` section, see below.

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 given that they all proxy for each other.

Finally, let's take a look at the genre columns, to get a sense of the relative frequencies:



Methods and Results

The model used here is ordinary least squares linear regression, with profit modeled as a linear combination of all the predictors:

$profit = \beta_0 + \beta_1 \cdot release\_date + \beta_2 \cdot budget + \beta_3 \cdot average\_rating + \sum_i \beta_i \cdot genre_i + \varepsilon$

Genres are treated as indicator variables with `TRUE` corresponding to `1` (more specifically, R casts them to factors with `FALSE` being the reference point). `release_date` is a `date` type, which is treated as an integer in linear regression (i.e. the number of days since `1970-01-01`).

After the above model is fit, we call `R`'s `step` function, which iteratively drops variables by testing if removing them from the model lowers the `AIC`.

Complete Model

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
```

Reduced Form Model

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

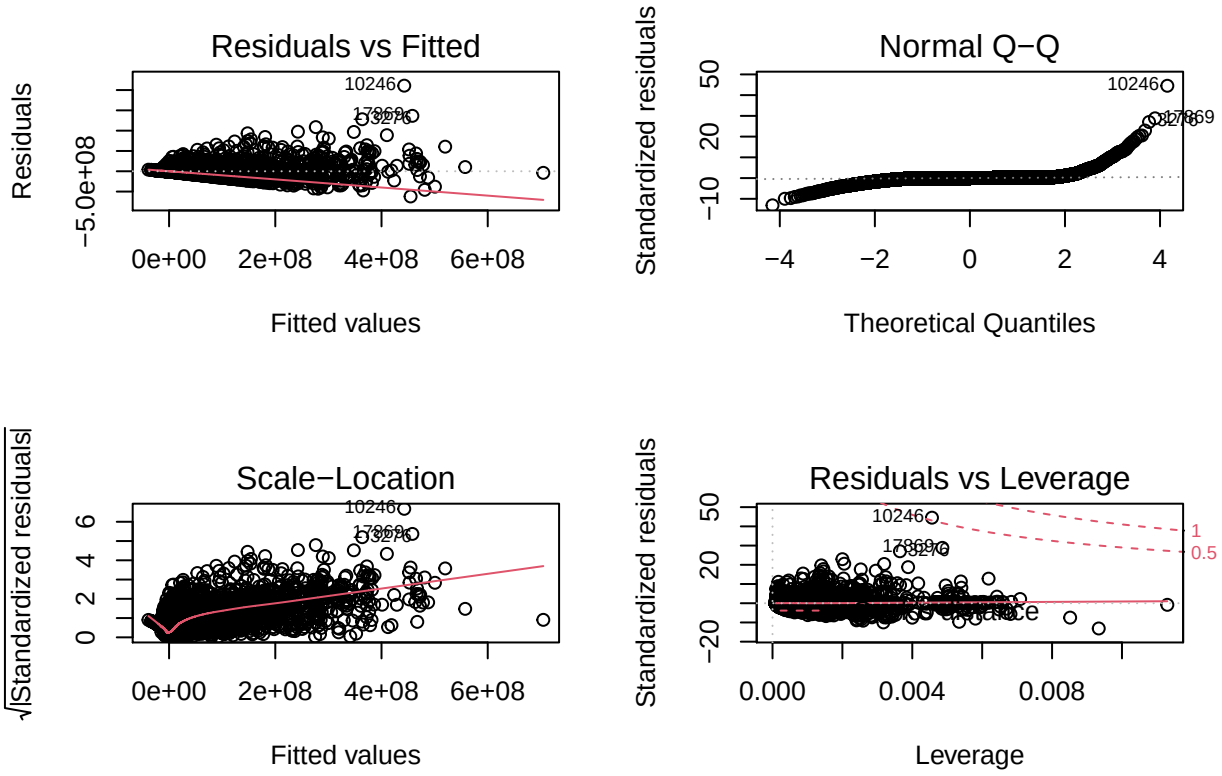
```
##                         Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)         -1.3909e+07  1.4602e+06  -9.5253 < 2.2e-16
## budget               1.8421e+00  1.4041e-02 131.1933 < 2.2e-16
## release_date         8.6901e+01  6.0599e+01   1.4340 0.1515726
## average_rating       4.0986e+06  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.

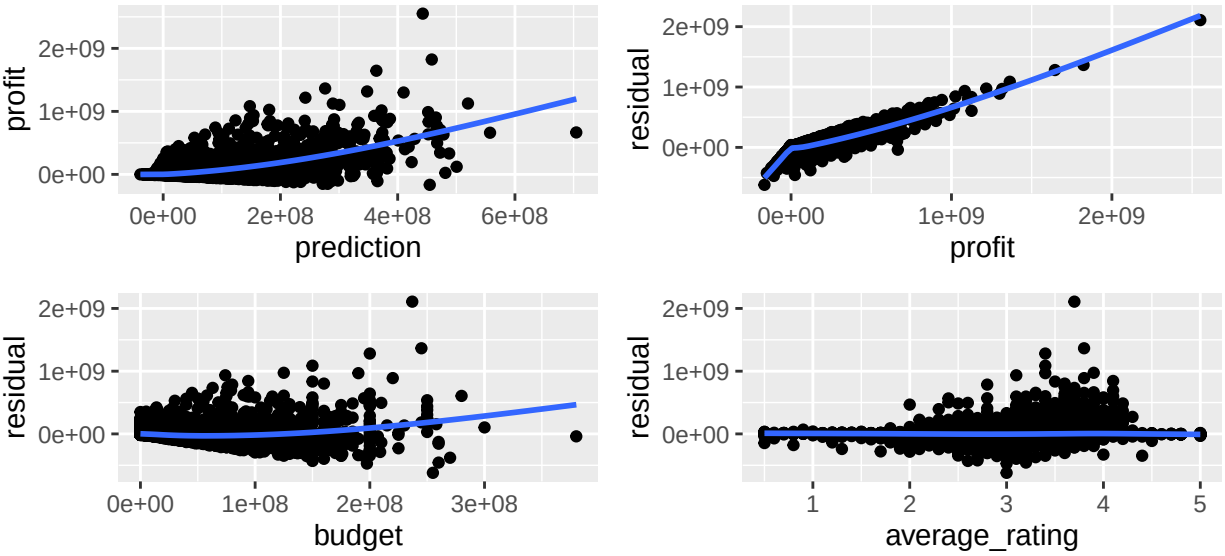
Diagnostics

Shown below are several diagnostic plots – note in particular the QQ plot of the residuals, and residuals vs. fitted plot:



Here are plots of the residuals against some of the predictors:

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Discussion

Model formulation

Ordinary least squares regression was chosen for this model because the response variable `profit` is continuous and its range is (in principle) the whole real line, with some films being profit negative.

The `step` function was chosen as the primary technique for this model so as 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

Model Coefficients: Genre, Release Date, Budget, Rating

The AIC minimization procedure eliminated 7 variables from the complete model, all of which are genres. The general trend is that niche genres such as `tv_movie` are removed in favor of mainstream ones such as `action`. Those that remain have effects on the order of `10e6` – a change in genre can make a difference on the order of `$1,000,000` in profit.

Given a threshold of .05, all variables in the reduced model are significant with the exceptions of `release_date` and three genres – `fantasy`, `family`, and `foreign`. See below for 95% confidence intervals of the coefficients. For the variables that pass significance, the confidence intervals estimate a range that is on the order of magnitude of the estimated coefficient:

	2.5 %	97.5 %
(Intercept)	-1.677107e+07	-1.104689e+07
budget	1.814555e+00	1.869597e+00
release_date	-3.187554e+01	2.056779e+02
average_rating	3.354968e+06	4.842332e+06
genre_thrillerTRUE	-4.898454e+06	-1.826815e+06
genre_warTRUE	-1.102010e+07	-2.953016e+06
genre_fantasyTRUE	-6.569835e+05	4.390114e+06
genre_actionTRUE	-5.228350e+06	-1.845547e+06
genre_familyTRUE	-6.309703e+05	4.059431e+06
genre_crimeTRUE	-4.472663e+06	-3.842064e+05
genre_foreignTRUE	-3.457126e+05	4.996523e+06
genre_adventureTRUE	3.024216e+06	7.739597e+06
genre_horrorTRUE	1.150590e+06	5.019758e+06
genre_dramaTRUE	-3.407863e+06	-1.140264e+06
genre_animationTRUE	1.575696e+06	6.834171e+06
genre_westernTRUE	-2.293264e+07	-1.029895e+07
genre_historyTRUE	-1.135554e+07	-4.628697e+06

The only significant genres which have positive coefficients are `adventure`, `horror`, and `animation`. Genres `action`, `crime`, `drama`, `war`, `thriller`, `history` and `western` all have negative coefficients. Taken together, these results suggest that audiences are averse to films with stressful subject matter (e.g. `drama` and `war`) or historical content (e.g. `history` and `western`). `horror` is a counterexample to this conclusion, but it is difficult to ascertain why this is the case without further investigation into how genre labels are assigned.

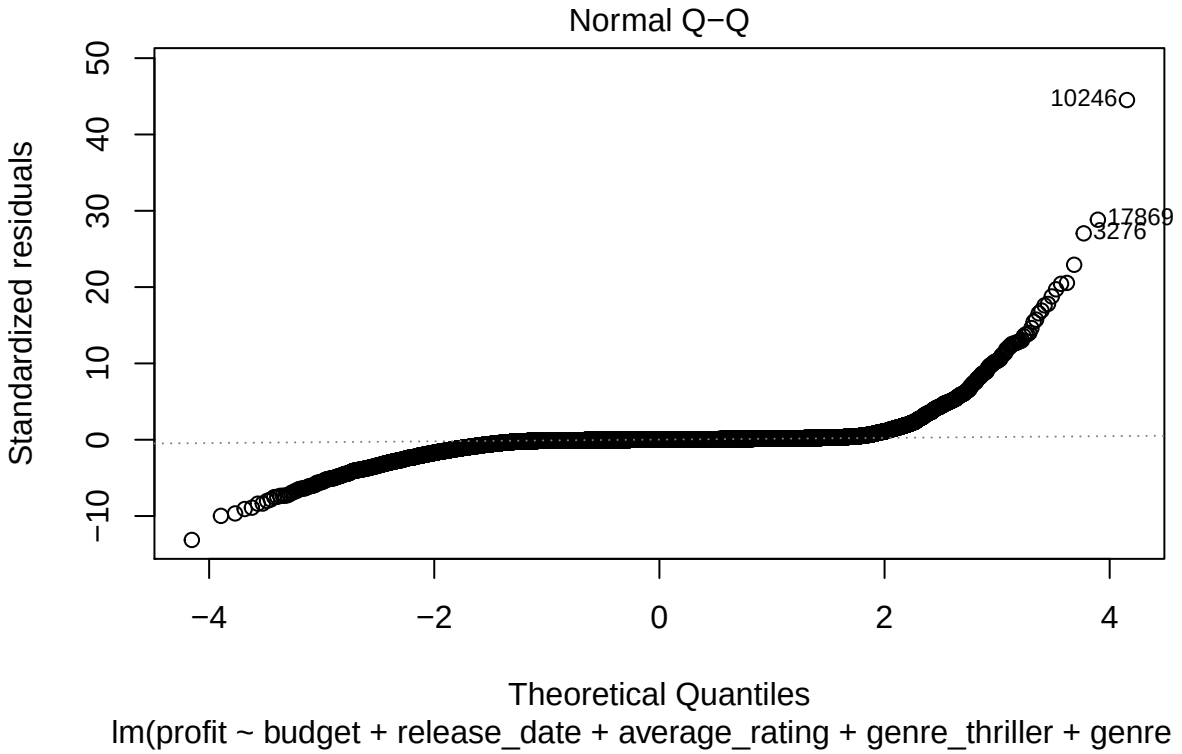
The remaining significant variables `budget` and `average_rating` have positive effects on `profit`. `budget` has a coefficient of `~1.8`, indicating a roughly 2:1 return for an added dollar of budget. `average_rating`

has a particularly strong effect on `profit`, with a coefficient of `4.0986e+06`; a single star increase in rating entails an increase in profit of $4,000,000. The effect of `release_date` is unclear, as its coefficient did not pass significance; the confidence interval of `release_date` crosses zero, and so the sign of this coefficient is unclear as well.

Altogether, the model suggests that the best way to make a profitable film is to create a lighthearted adventure film which is also a critical favorite, such as Harry Potter or Star Wars.

Diagnostics

Of the diagnostic graphs in the results section, the QQ plot and the residuals vs fitted plot indicate issues with the model. From the QQ plot, we see that the model errors are "heavy-tailed" rather than normally distributed.



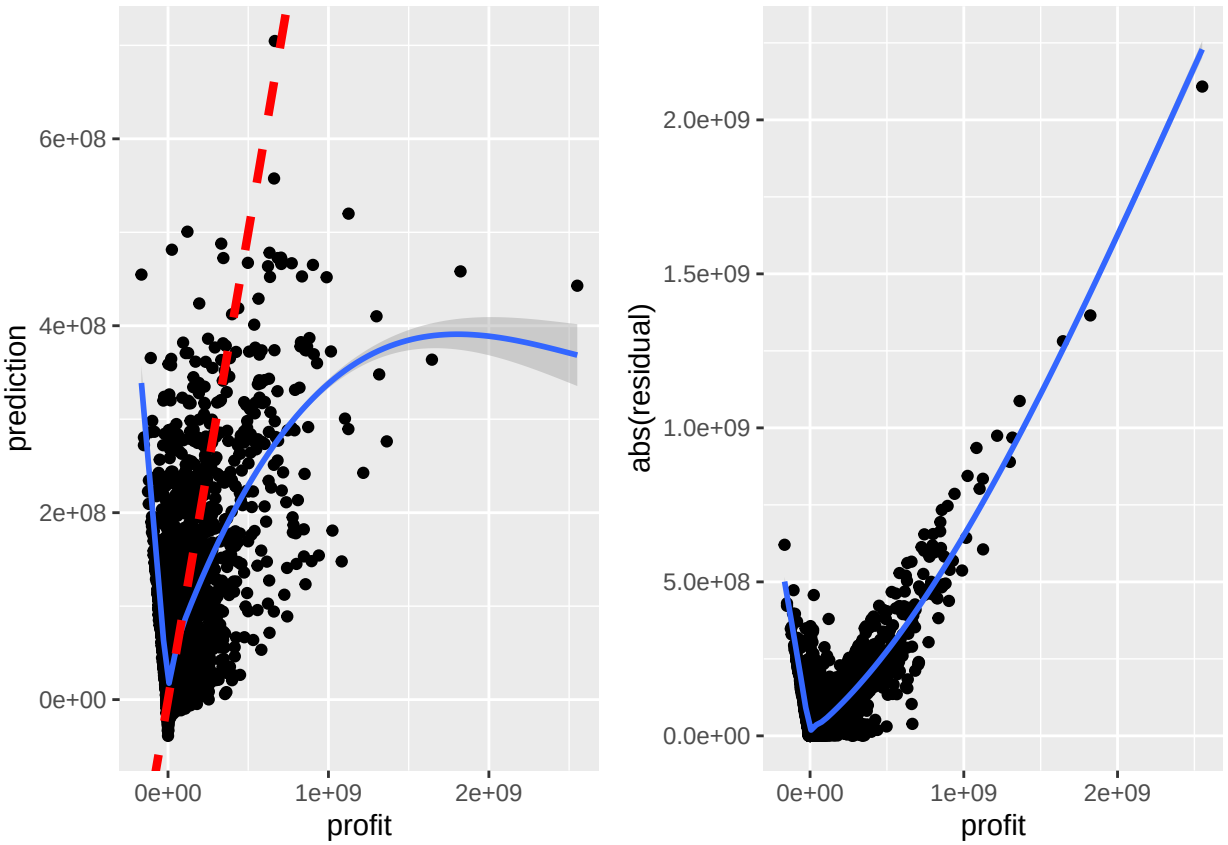
Though this has no impact on the sign and magnitude of our estimated coefficients, it does make our significance estimates and confidence intervals suspect. Heavy tailed residuals cause us to overestimate the variance, and so our confidence intervals are likely wider than they should be.

Second, we see from the fitted vs residuals (and the residuals vs predictors) plots that the residuals are still very much correlated with the prediction, response, and inputs after the regression. In fact, the plots of the residuals against the predictors closely resemble the plots of profit against the predictors.

The two problems above are due to the same underlying issue with the data – the response is heavily peaked at zero, and so the regression has trouble picking up signal because it is largely encouraged to skew predictions towards zero.

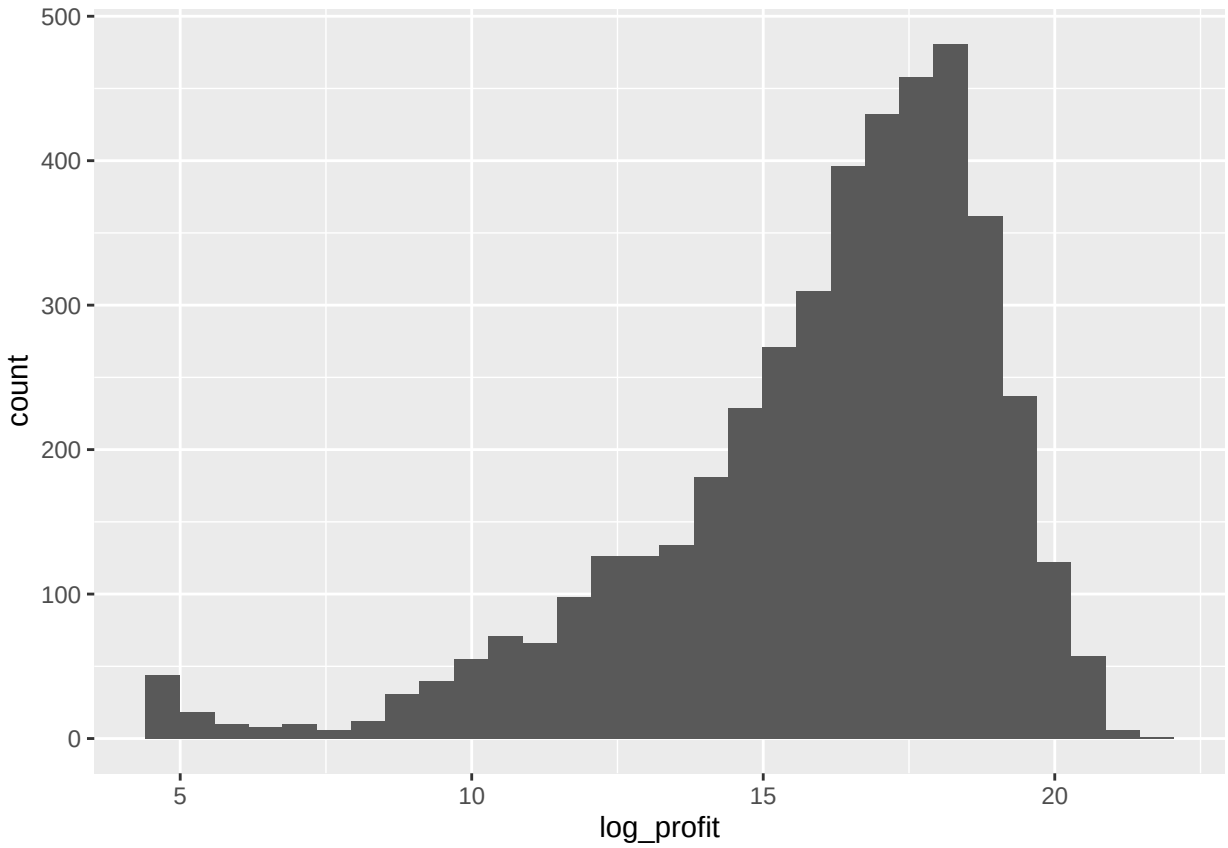
This is confirmed by the plots below. The graph on the left indicates that the model's predictions, especially for highly profitable films, fall closer to zero than expected. The graph on the right indicates that the absolute value of the residual increases as profit moves away from zero in either direction.

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



I attempted to solve this problem by restricting the analysis to films which were profit positive and log-transforming the `profit` variable. 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:

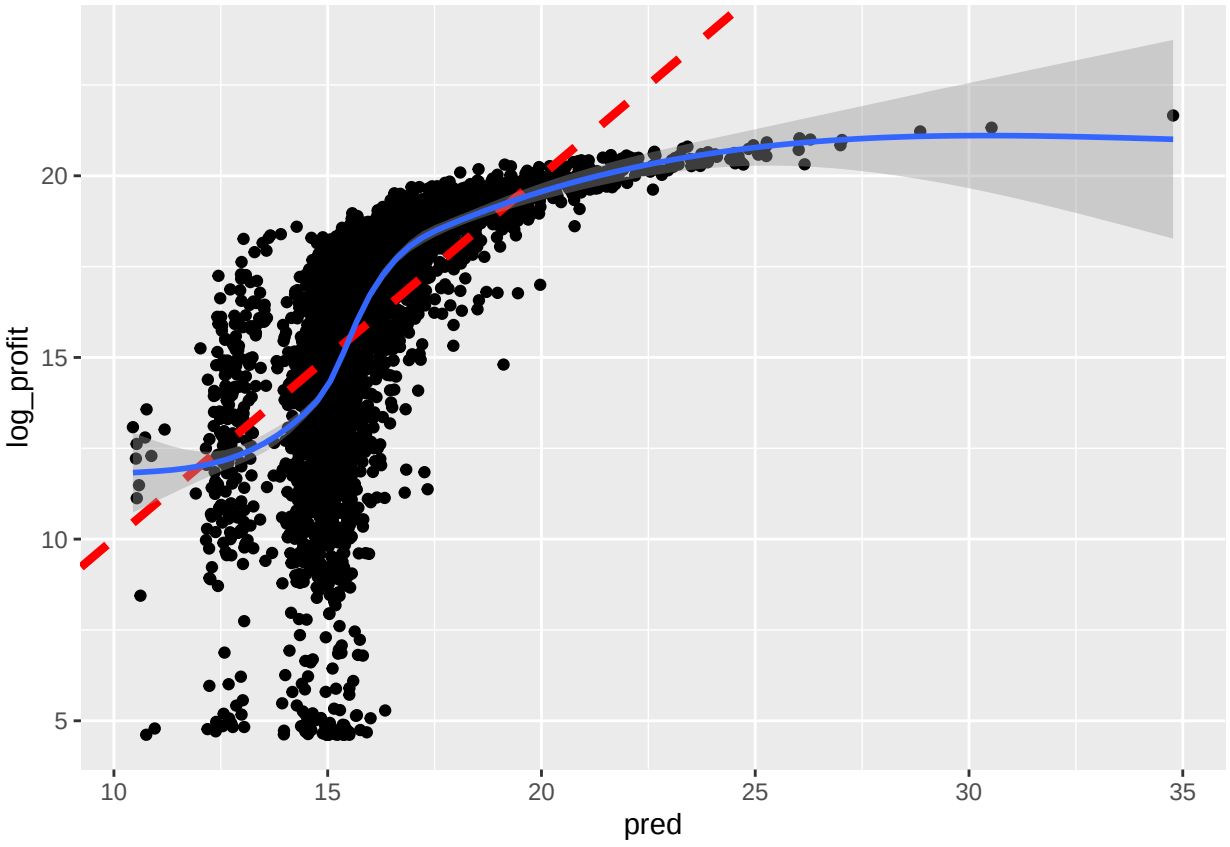
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Now let us run the same procedure as above, but with the response as `log(profit)`. This seems to improve things a bit for most points, but on the edges of the distribution the problem is now reversed; for highly profitable films, predictions are skewed high. A moderately different set of predictors and coefficients is determined under this model:

```
##                           Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)             1.4120e+01  3.1889e-01  44.2795 < 2.2e-16
## profit                  5.9194e-09  3.4147e-10  17.3353 < 2.2e-16
## budget                  1.9334e-08  1.2116e-09  15.9580 < 2.2e-16
## release_date           -6.4137e-05  1.3434e-05  -4.7740 1.865e-06
## average_rating          4.7628e-01  7.7917e-02   6.1126 1.065e-09
## genre_thrillerTRUE      2.4454e-01  1.0212e-01   2.3947 0.0166743
## genre_warTRUE           4.2227e-01  2.4147e-01   1.7488 0.0804025
## genre_documentaryTRUE -2.2910e+00  2.0070e-01 -11.4154 < 2.2e-16
## genre_actionTRUE        1.4480e-01  1.0100e-01   1.4337 0.1517263
## genre_familyTRUE        6.4429e-01  1.3096e-01   4.9197 8.986e-07
## genre_crimeTRUE         4.0507e-01  1.1391e-01   3.5559 0.0003806
## genre_comedyTRUE        3.9995e-01  9.0231e-02   4.4325 9.542e-06
## genre_foreignTRUE      -1.9951e+00  3.0951e-01  -6.4460 1.273e-10
## genre_mysteryTRUE       4.3756e-01  1.5133e-01   2.8913 0.0038550
## genre_horrorTRUE        5.7994e-01  1.4426e-01   4.0200 5.918e-05
## genre_dramaTRUE        -3.3964e-01  9.0853e-02  -3.7384 0.0001876
## genre_romanceTRUE       2.6344e-01  1.0033e-01   2.6257 0.0086765
## genre_historyTRUE       4.0367e-01  2.1015e-01   1.9208 0.0548159
## genre_musicTRUE         3.3636e-01  2.1551e-01   1.5607 0.1186601
##
## n = 4398, p = 19, Residual SE = 2.37981, R-Squared = 0.38
```

```
## [1] AIC for model with log transformation: 20128.247262
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Future Directions and Improvements

As discussed above, the primary affliction of this model is the skew of the response towards zero. The main improvement that should be done here is to change the link function to account for heavy tails. Perhaps robust regression would be a suitable technique here, or a link function that models an exponential response rather than a gaussian.

A second improvement would be a wider diversity of input variables to the model; genre may be correlated with numerous external variables that are not controlled for (e.g. marketing or the reputation of particular directors) and so the direct effect of a genre is somewhat dubious from this investigation. Further investigation might also try to extract features from the "ground truth" e.g. film scripts or images.