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

Profitability has become a contentious subject in the film industry in recent years. Acclaimed director G. Inarritu likened the state of affairs of the industry "the 99% and the 1%" of the global economy (https://youtu.be/SQ7qKKQrSBY), with the highest grossing films being either remakes of one kind or another (e.g. Marvel, Star Wars), or very small-budget independent films that receive high critical praise (e.g. Moonlight, Birdman, etc.).

The goal of this paper is to empirically determine predictors of the profitability of a film using Kaggle's "The Movies Dataset", an aggregate of data from GroupLens and TMDB, containing data on film revenue, budget, genre, and ratings.

We begin from a complete set of predictors (see below) and use an iterative procedure to minimize the Akaike-Information Criterion (i.e. R's `step` function) in order to discover the most relevant variables.

- Release date. This allows us capture some kind of time-trend in profitability
- Budget. This captures the constraint that movies must be generally profitable.
- Genre. This captures popularity effects by genre.
- Rating. This captures the critical-reception of the film.

Methods

Data

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,
  budged: float,
  imdbId: int
}

RATINGS {
  movieId: int,
  userId: int,
```

```
  rating: float,
}

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

Data Transformation and Cleaning

Some preparation was needed to turn these three tables into a usable format. The genre column of `movies_metadata` is JSON, encoded as a list of pairs representing the set of genres the film is associated with e.g. `[{genreId: 1, genreName: Action}, {genreId: 2, genreName: Comedy}]`; this column was converted to a set of boolean columns e.g. `genre_action` with `True` representing that the film belongs to that genre.

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. Finally, a join was performed between the two tables using the join table `links` to create a single `movies` table, which serves as the final dataset on which we run our regression.

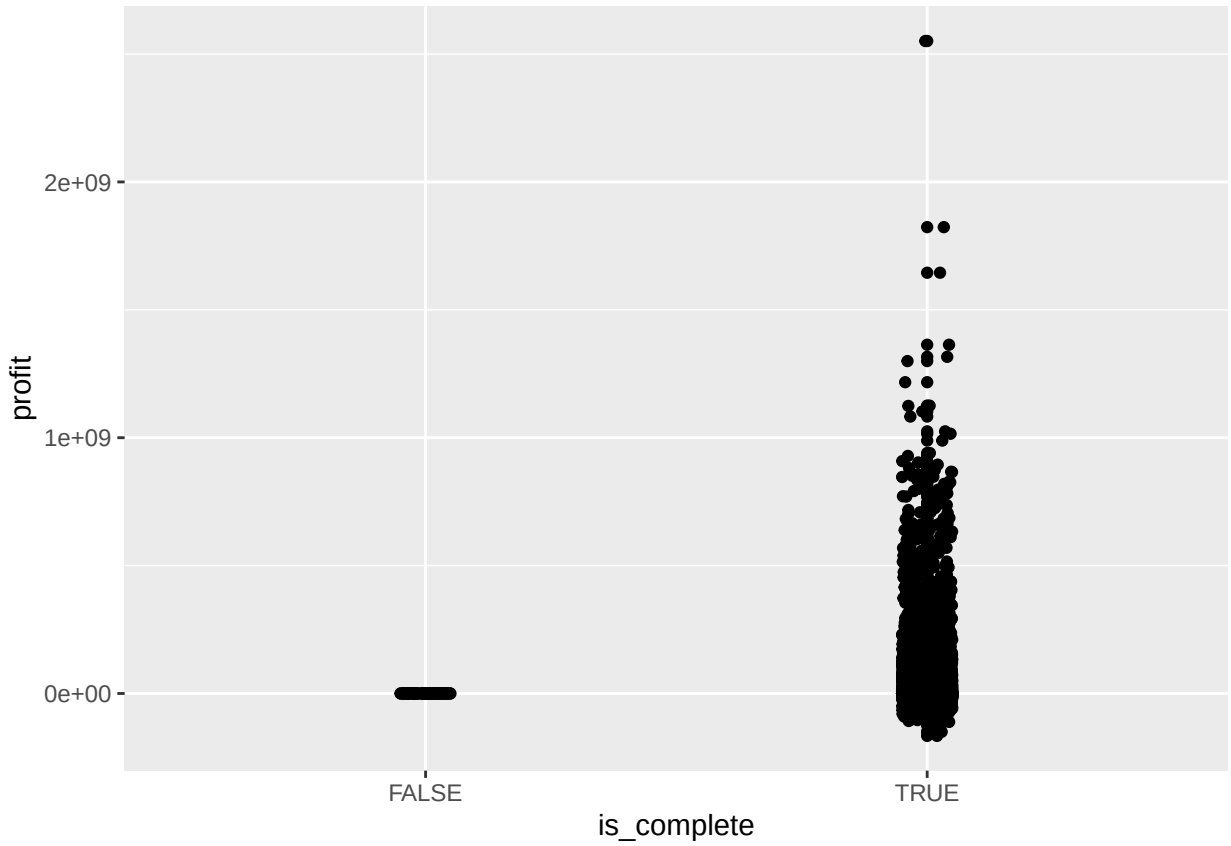
The `movies` schema is described below. The `profit`column is a derived quantity defined as `revenue - budget`; note the assumption here that all films use exactly there budget. The table also contains 32 genre columns with naming pattern `genre_[name]`; the prefix of "genre_" helps to avoid unfortunate name conflicts with keywords in postgres and python e.g. foreign (as in foreign key).

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

Finally, rows with missing values are dropped. As can be seen from the chart below, the variable `is_complete` may be correlated with revenue, with incomplete rows typically having `0` profit, but this is not a major conceptual issue for the project for two reasons; first, the vast majority of the films in the dataset are roughly break even.

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

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



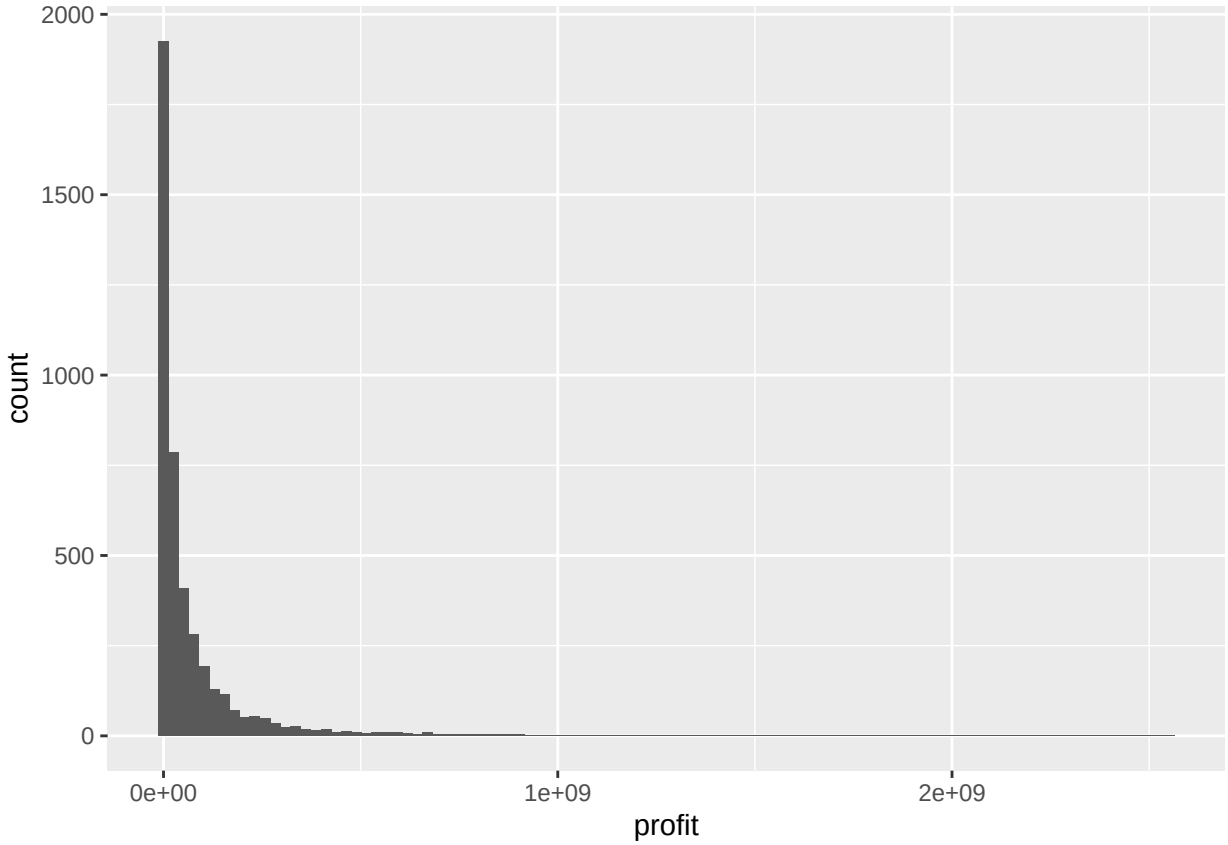

```
## [1] 30615
```

```
## [1] 4344
```

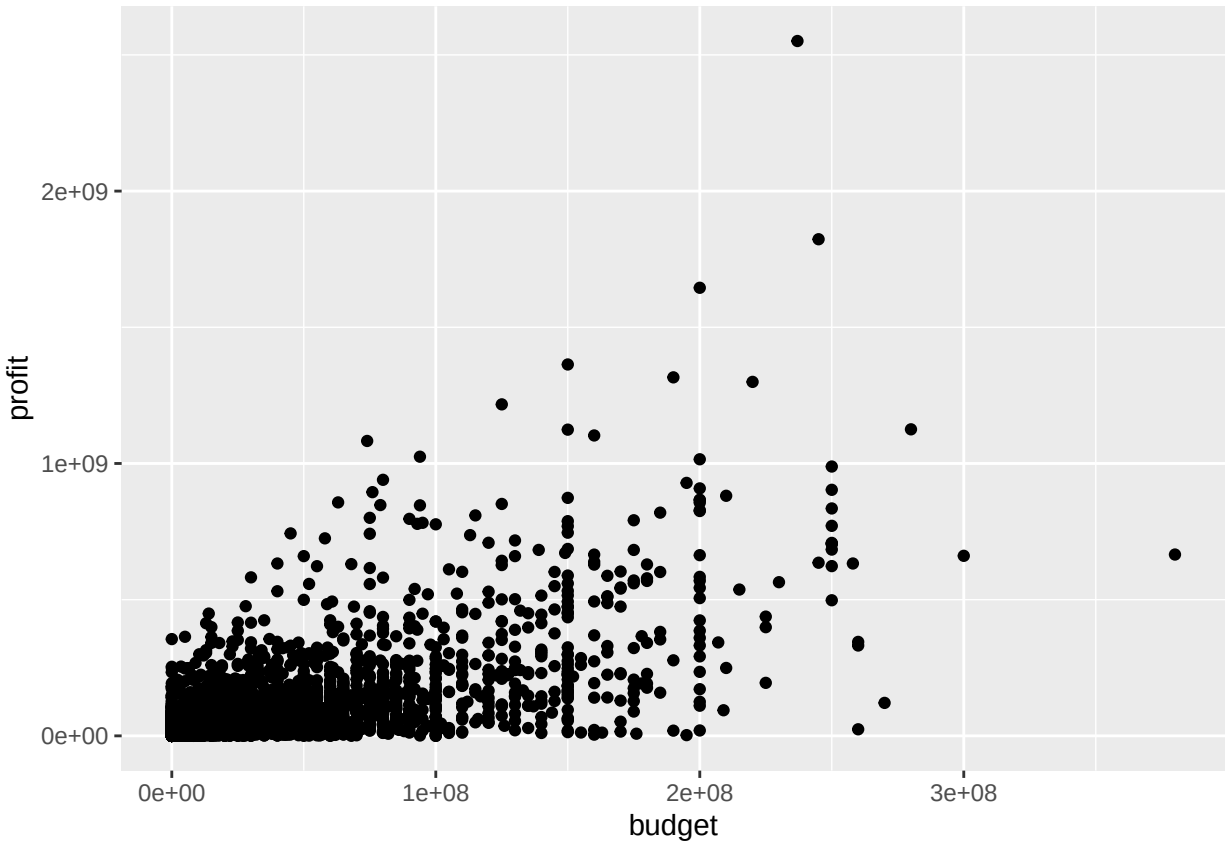
General Characteristics and Distribution

First, let's take a look at the distribution of profit. It seems profit is approximately a gaussian variable, heavily peaked at zero, with a fraction of films being either profit positive or profit negative:

```
## Warning: Removed 3 rows containing non-finite values (stat_bin).
```

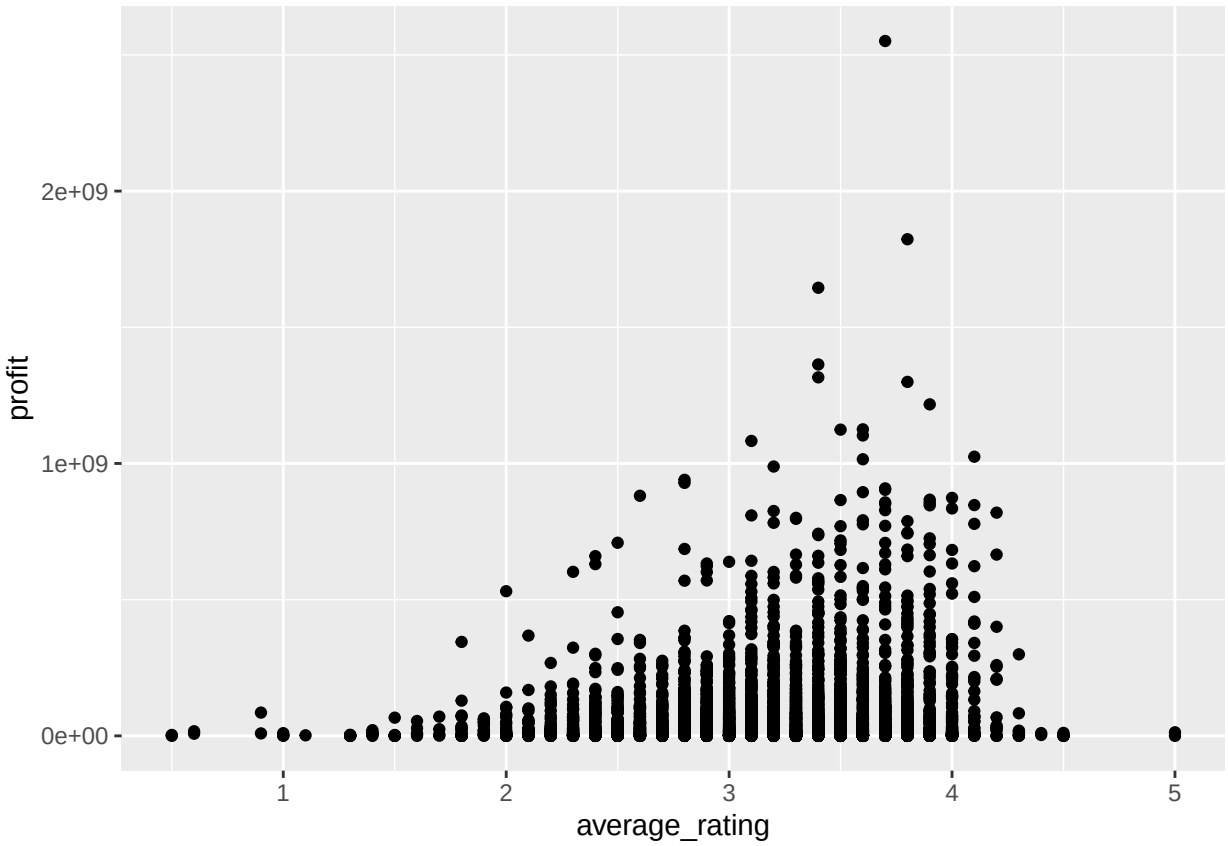



```
## Warning: Removed 3 rows containing missing values (geom_point).
```



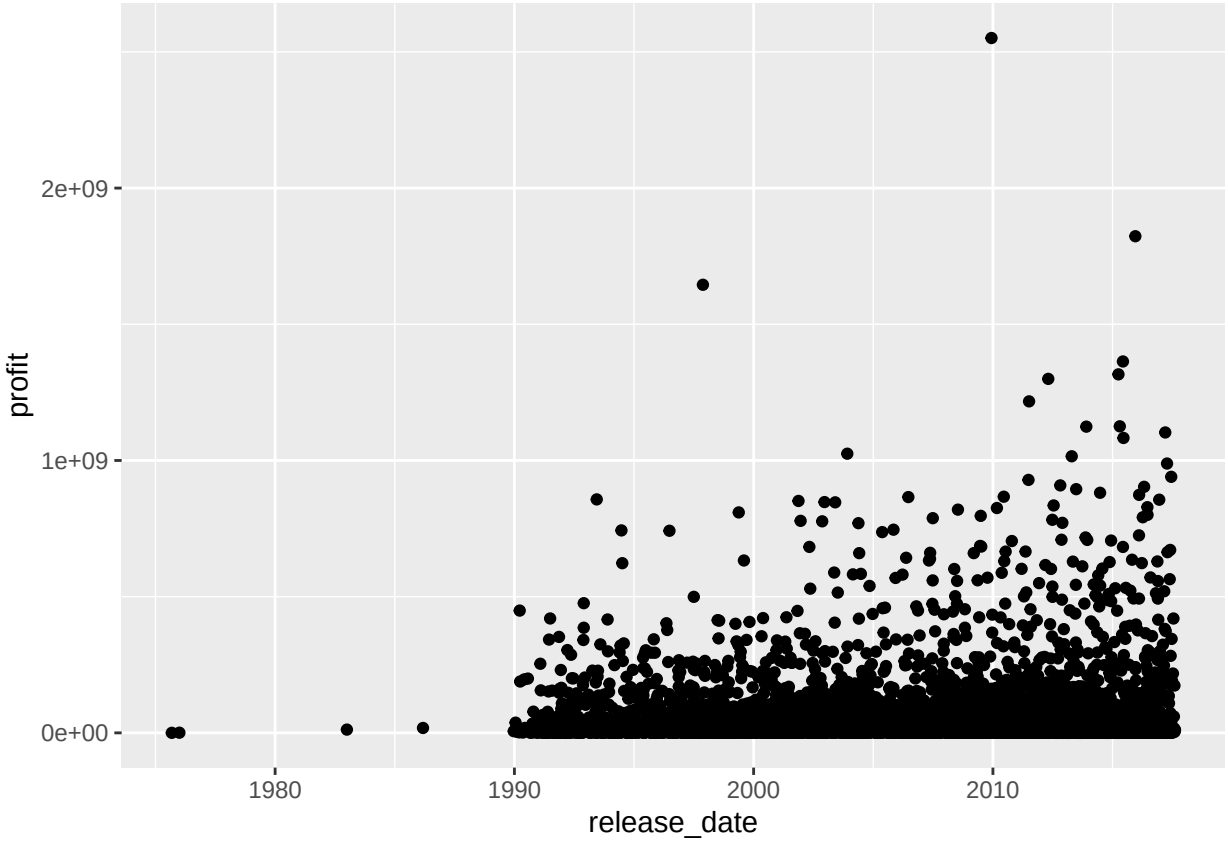
Ratings appear to have some effect on profitability, but as we can clearly see the relationship is non-linear:

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



Release date has an interesting relationship with profit;

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



Model

```
## Start:  AIC=160429.1
## profit ~ budget + average_rating + release_date + genre_vision_view_entertainment +
##     genre_thriller + genre_pulser_productions + genre_war + genre_fantasy +
##     genre_documentary + genre_sentai_filmworks + genre_gohands +
##     genre_action + genre_mardock_scramble_production_committee +
##     genre_family + genre_crime + genre_comedy + genre_foreign +
##     genre_adventure + genre_telescene_film_group_productions +
##     genre_mystery + genre_science_fiction + genre_horror + genre_drama +
##     genre_carousel_productions + genre_rogue_state + genre_the_cartel +
##     genre_animation + genre_romance + genre_western + genre_aniplex +
##     genre_history + genre_brosta_tv + genre_tv_movie + genre_odyssey_media +
##     genre_music
##
##
## Step:  AIC=160429.1
## profit ~ budget + average_rating + release_date + genre_vision_view_entertainment +
##     genre_thriller + genre_pulser_productions + genre_war + genre_fantasy +
##     genre_documentary + genre_sentai_filmworks + genre_gohands +
##     genre_action + genre_mardock_scramble_production_committee +
##     genre_family + genre_crime + genre_comedy + genre_foreign +
##     genre_adventure + genre_telescene_film_group_productions +
##     genre_mystery + genre_science_fiction + genre_horror + genre_drama +
##     genre_carousel_productions + genre_rogue_state + genre_the_cartel +
##     genre_animation + genre_romance + genre_western + genre_aniplex +
##     genre_history + genre_brosta_tv + genre_tv_movie + genre_music
```

```
##
##
## Step:  AIC=160429.1
## profit ~ budget + average_rating + release_date + genre_vision_view_entertainment +
##     genre_thriller + genre_pulser_productions + genre_war + genre_fantasy +
##     genre_documentary + genre_sentai_filmworks + genre_gohands +
##     genre_action + genre_mardock_scramble_production_committee +
##     genre_family + genre_crime + genre_comedy + genre_foreign +
##     genre_adventure + genre_telescene_film_group_productions +
##     genre_mystery + genre_science_fiction + genre_horror + genre_drama +
##     genre_carousel_productions + genre_rogue_state + genre_the_cartel +
##     genre_animation + genre_romance + genre_western + genre_aniplex +
##     genre_history + genre_tv_movie + genre_music
##
##
## Step:  AIC=160429.1
## profit ~ budget + average_rating + release_date + genre_vision_view_entertainment +
##     genre_thriller + genre_pulser_productions + genre_war + genre_fantasy +
##     genre_documentary + genre_sentai_filmworks + genre_gohands +
##     genre_action + genre_mardock_scramble_production_committee +
##     genre_family + genre_crime + genre_comedy + genre_foreign +
##     genre_adventure + genre_telescene_film_group_productions +
##     genre_mystery + genre_science_fiction + genre_horror + genre_drama +
##     genre_carousel_productions + genre_rogue_state + genre_the_cartel +
##     genre_animation + genre_romance + genre_western + genre_history +
##     genre_tv_movie + genre_music
##
##
## Step:  AIC=160429.1
## profit ~ budget + average_rating + release_date + genre_vision_view_entertainment +
##     genre_thriller + genre_pulser_productions + genre_war + genre_fantasy +
##     genre_documentary + genre_sentai_filmworks + genre_gohands +
##     genre_action + genre_mardock_scramble_production_committee +
##     genre_family + genre_crime + genre_comedy + genre_foreign +
##     genre_adventure + genre_telescene_film_group_productions +
##     genre_mystery + genre_science_fiction + genre_horror + genre_drama +
##     genre_carousel_productions + genre_rogue_state + genre_animation +
##     genre_romance + genre_western + genre_history + genre_tv_movie +
##     genre_music
##
##
## Step:  AIC=160429.1
## profit ~ budget + average_rating + release_date + genre_vision_view_entertainment +
##     genre_thriller + genre_pulser_productions + genre_war + genre_fantasy +
##     genre_documentary + genre_sentai_filmworks + genre_gohands +
##     genre_action + genre_mardock_scramble_production_committee +
##     genre_family + genre_crime + genre_comedy + genre_foreign +
##     genre_adventure + genre_telescene_film_group_productions +
##     genre_mystery + genre_science_fiction + genre_horror + genre_drama +
##     genre_carousel_productions + genre_animation + genre_romance +
##     genre_western + genre_history + genre_tv_movie + genre_music
##
##
## Step:  AIC=160429.1
```

```
## profit ~ budget + average_rating + release_date + genre_vision_view_entertainment +
##     genre_thriller + genre_pulser_productions + genre_war + genre_fantasy +
##     genre_documentary + genre_sentai_filmworks + genre_gohands +
##     genre_action + genre_mardock_scramble_production_committee +
##     genre_family + genre_crime + genre_comedy + genre_foreign +
##     genre_adventure + genre_telescene_film_group_productions +
##     genre_mystery + genre_science_fiction + genre_horror + genre_drama +
##     genre_animation + genre_romance + genre_western + genre_history +
##     genre_tv_movie + genre_music
##
##
## Step:  AIC=160429.1
## profit ~ budget + average_rating + release_date + genre_vision_view_entertainment +
##     genre_thriller + genre_pulser_productions + genre_war + genre_fantasy +
##     genre_documentary + genre_sentai_filmworks + genre_gohands +
##     genre_action + genre_mardock_scramble_production_committee +
##     genre_family + genre_crime + genre_comedy + genre_foreign +
##     genre_adventure + genre_mystery + genre_science_fiction +
##     genre_horror + genre_drama + genre_animation + genre_romance +
##     genre_western + genre_history + genre_tv_movie + genre_music
##
##
## Step:  AIC=160429.1
## profit ~ budget + average_rating + release_date + genre_vision_view_entertainment +
##     genre_thriller + genre_pulser_productions + genre_war + genre_fantasy +
##     genre_documentary + genre_sentai_filmworks + genre_gohands +
##     genre_action + genre_family + genre_crime + genre_comedy +
##     genre_foreign + genre_adventure + genre_mystery + genre_science_fiction +
##     genre_horror + genre_drama + genre_animation + genre_romance +
##     genre_western + genre_history + genre_tv_movie + genre_music
##
##
## Step:  AIC=160429.1
## profit ~ budget + average_rating + release_date + genre_vision_view_entertainment +
##     genre_thriller + genre_pulser_productions + genre_war + genre_fantasy +
##     genre_documentary + genre_sentai_filmworks + genre_action +
##     genre_family + genre_crime + genre_comedy + genre_foreign +
##     genre_adventure + genre_mystery + genre_science_fiction +
##     genre_horror + genre_drama + genre_animation + genre_romance +
##     genre_western + genre_history + genre_tv_movie + genre_music
##
##
## Step:  AIC=160429.1
## profit ~ budget + average_rating + release_date + genre_vision_view_entertainment +
##     genre_thriller + genre_pulser_productions + genre_war + genre_fantasy +
##     genre_documentary + genre_action + genre_family + genre_crime +
##     genre_comedy + genre_foreign + genre_adventure + genre_mystery +
##     genre_science_fiction + genre_horror + genre_drama + genre_animation +
##     genre_romance + genre_western + genre_history + genre_tv_movie +
##     genre_music
##
##
## Step:  AIC=160429.1
## profit ~ budget + average_rating + release_date + genre_vision_view_entertainment +
```

```
##     genre_thriller + genre_war + genre_fantasy + genre_documentary +
##     genre_action + genre_family + genre_crime + genre_comedy +
##     genre_foreign + genre_adventure + genre_mystery + genre_science_fiction +
##     genre_horror + genre_drama + genre_animation + genre_romance +
##     genre_western + genre_history + genre_tv_movie + genre_music
##
##
## Step:  AIC=160429.1
## profit ~ budget + average_rating + release_date + genre_thriller +
##     genre_war + genre_fantasy + genre_documentary + genre_action +
##     genre_family + genre_crime + genre_comedy + genre_foreign +
##     genre_adventure + genre_mystery + genre_science_fiction +
##     genre_horror + genre_drama + genre_animation + genre_romance +
##     genre_western + genre_history + genre_tv_movie + genre_music
##
##                         Df  Sum of Sq        RSS    AIC
## - genre_tv_movie         1 3.0125e+13 4.8182e+19 160427
## - genre_foreign          1 2.0755e+14 4.8182e+19 160427
## - genre_comedy           1 1.0604e+15 4.8183e+19 160427
## - release_date           1 1.2711e+15 4.8183e+19 160427
## - genre_mystery          1 2.4918e+15 4.8184e+19 160427
## - genre_music            1 1.3031e+16 4.8195e+19 160428
## - genre_fantasy          1 1.3361e+16 4.8195e+19 160428
## - genre_action           1 1.8718e+16 4.8201e+19 160429
## <none>                                4.8182e+19 160429
## - genre_war              1 2.2626e+16 4.8204e+19 160429
## - genre_crime            1 2.2919e+16 4.8205e+19 160429
## - genre_science_fiction  1 2.6402e+16 4.8208e+19 160429
## - genre_family           1 3.0855e+16 4.8213e+19 160430
## - genre_romance          1 3.2689e+16 4.8215e+19 160430
## - genre_animation        1 3.6451e+16 4.8218e+19 160430
## - genre_thriller         1 3.8304e+16 4.8220e+19 160431
## - genre_history          1 5.4131e+16 4.8236e+19 160432
## - genre_horror           1 6.0519e+16 4.8242e+19 160433
## - genre_documentary      1 8.3562e+16 4.8265e+19 160435
## - genre_western          1 1.0843e+17 4.8290e+19 160437
## - genre_drama            1 1.3942e+17 4.8321e+19 160440
## - genre_adventure        1 1.7367e+17 4.8355e+19 160443
## - average_rating         1 1.3954e+18 4.9577e+19 160551
## - budget                 1 2.5237e+19 7.3419e+19 162255
##
## Step:  AIC=160427.1
## profit ~ budget + average_rating + release_date + genre_thriller +
##     genre_war + genre_fantasy + genre_documentary + genre_action +
##     genre_family + genre_crime + genre_comedy + genre_foreign +
##     genre_adventure + genre_mystery + genre_science_fiction +
##     genre_horror + genre_drama + genre_animation + genre_romance +
##     genre_western + genre_history + genre_music
##
##                         Df  Sum of Sq        RSS    AIC
## - genre_foreign          1 2.0786e+14 4.8182e+19 160425
## - genre_comedy           1 1.0622e+15 4.8183e+19 160425
## - release_date           1 1.2798e+15 4.8183e+19 160425
## - genre_mystery          1 2.4945e+15 4.8184e+19 160425
```

```
## - genre_music            1 1.3030e+16 4.8195e+19 160426
## - genre_fantasy          1 1.3350e+16 4.8195e+19 160426
## - genre_action           1 1.8765e+16 4.8201e+19 160427
## <none>                                4.8182e+19 160427
## - genre_war              1 2.2627e+16 4.8204e+19 160427
## - genre_crime            1 2.2913e+16 4.8205e+19 160427
## - genre_science_fiction  1 2.6376e+16 4.8208e+19 160427
## - genre_family           1 3.0841e+16 4.8213e+19 160428
## - genre_romance          1 3.2670e+16 4.8215e+19 160428
## - genre_animation        1 3.6438e+16 4.8218e+19 160428
## - genre_thriller         1 3.8330e+16 4.8220e+19 160429
## - genre_history          1 5.4143e+16 4.8236e+19 160430
## - genre_horror           1 6.0511e+16 4.8242e+19 160431
## - genre_documentary      1 8.3569e+16 4.8265e+19 160433
## - genre_western          1 1.0845e+17 4.8290e+19 160435
## - genre_drama            1 1.3939e+17 4.8321e+19 160438
## - genre_adventure        1 1.7423e+17 4.8356e+19 160441
## - average_rating         1 1.3954e+18 4.9577e+19 160549
## - budget                 1 2.5247e+19 7.3429e+19 162254
##
## Step:  AIC=160425.1
## profit ~ budget + average_rating + release_date + genre_thriller +
##     genre_war + genre_fantasy + genre_documentary + genre_action +
##     genre_family + genre_crime + genre_comedy + genre_adventure +
##     genre_mystery + genre_science_fiction + genre_horror + genre_drama +
##     genre_animation + genre_romance + genre_western + genre_history +
##     genre_music
##
##                         Df  Sum of Sq        RSS    AIC
## - genre_comedy           1 1.0492e+15 4.8183e+19 160423
## - release_date           1 1.2554e+15 4.8183e+19 160423
## - genre_mystery          1 2.5033e+15 4.8185e+19 160423
## - genre_music            1 1.3150e+16 4.8195e+19 160424
## - genre_fantasy          1 1.3382e+16 4.8195e+19 160424
## - genre_action           1 1.8886e+16 4.8201e+19 160425
## <none>                                4.8182e+19 160425
## - genre_war              1 2.2545e+16 4.8205e+19 160425
## - genre_crime            1 2.2818e+16 4.8205e+19 160425
## - genre_science_fiction  1 2.6303e+16 4.8208e+19 160425
## - genre_family           1 3.0895e+16 4.8213e+19 160426
## - genre_romance          1 3.2570e+16 4.8215e+19 160426
## - genre_animation        1 3.6452e+16 4.8219e+19 160426
## - genre_thriller         1 3.8217e+16 4.8220e+19 160427
## - genre_history          1 5.4147e+16 4.8236e+19 160428
## - genre_horror           1 6.0494e+16 4.8243e+19 160429
## - genre_documentary      1 8.4593e+16 4.8267e+19 160431
## - genre_western          1 1.0838e+17 4.8290e+19 160433
## - genre_drama            1 1.3962e+17 4.8322e+19 160436
## - genre_adventure        1 1.7430e+17 4.8356e+19 160439
## - average_rating         1 1.3974e+18 4.9579e+19 160547
## - budget                 1 2.5288e+19 7.3470e+19 162255
##
## Step:  AIC=160423.2
## profit ~ budget + average_rating + release_date + genre_thriller +
```

```
##     genre_war + genre_fantasy + genre_documentary + genre_action +
##     genre_family + genre_crime + genre_adventure + genre_mystery +
##     genre_science_fiction + genre_horror + genre_drama + genre_animation +
##     genre_romance + genre_western + genre_history + genre_music
##
##                          Df  Sum of Sq        RSS    AIC
## - release_date            1 1.1733e+15 4.8184e+19 160421
## - genre_mystery           1 2.2724e+15 4.8185e+19 160421
## - genre_music             1 1.3464e+16 4.8197e+19 160422
## - genre_fantasy           1 1.4046e+16 4.8197e+19 160422
## - genre_action            1 1.8165e+16 4.8201e+19 160423
## - genre_war               1 2.2056e+16 4.8205e+19 160423
## <none>                                 4.8183e+19 160423
## - genre_crime             1 2.2723e+16 4.8206e+19 160423
## - genre_science_fiction   1 2.5487e+16 4.8209e+19 160423
## - genre_family            1 3.0099e+16 4.8213e+19 160424
## - genre_romance           1 3.1711e+16 4.8215e+19 160424
## - genre_thriller          1 3.7654e+16 4.8221e+19 160425
## - genre_animation         1 3.8057e+16 4.8221e+19 160425
## - genre_history           1 5.3311e+16 4.8236e+19 160426
## - genre_horror            1 6.6311e+16 4.8249e+19 160427
## - genre_documentary       1 8.6131e+16 4.8269e+19 160429
## - genre_western           1 1.0801e+17 4.8291e+19 160431
## - genre_drama             1 1.5021e+17 4.8333e+19 160435
## - genre_adventure         1 1.7839e+17 4.8361e+19 160437
## - average_rating          1 1.4084e+18 4.9592e+19 160546
## - budget                  1 2.5402e+19 7.3586e+19 162259
##
## Step:  AIC=160421.3
## profit ~ budget + average_rating + genre_thriller + genre_war +
##     genre_fantasy + genre_documentary + genre_action + genre_family +
##     genre_crime + genre_adventure + genre_mystery + genre_science_fiction +
##     genre_horror + genre_drama + genre_animation + genre_romance +
##     genre_western + genre_history + genre_music
##
##                          Df  Sum of Sq        RSS    AIC
## - genre_mystery           1 2.2002e+15 4.8186e+19 160420
## - genre_music             1 1.3340e+16 4.8198e+19 160421
## - genre_fantasy           1 1.4224e+16 4.8198e+19 160421
## - genre_action            1 1.7916e+16 4.8202e+19 160421
## - genre_war               1 2.1988e+16 4.8206e+19 160421
## - genre_crime             1 2.2042e+16 4.8206e+19 160421
## <none>                                 4.8184e+19 160421
## - genre_science_fiction   1 2.5173e+16 4.8209e+19 160422
## - genre_family            1 3.1567e+16 4.8216e+19 160422
## - genre_romance           1 3.3394e+16 4.8218e+19 160422
## - genre_animation         1 3.7233e+16 4.8222e+19 160423
## - genre_thriller          1 3.7345e+16 4.8222e+19 160423
## - genre_history           1 5.3617e+16 4.8238e+19 160424
## - genre_horror            1 6.5969e+16 4.8250e+19 160425
## - genre_documentary       1 8.6841e+16 4.8271e+19 160427
## - genre_western           1 1.0806e+17 4.8292e+19 160429
## - genre_drama             1 1.5037e+17 4.8335e+19 160433
## - genre_adventure         1 1.7922e+17 4.8363e+19 160435
```

```
## - average_rating         1 1.4076e+18 4.9592e+19 160544
## - budget                 1 2.5730e+19 7.3914e+19 162277
##
## Step:  AIC=160419.5
## profit ~ budget + average_rating + genre_thriller + genre_war +
##     genre_fantasy + genre_documentary + genre_action + genre_family +
##     genre_crime + genre_adventure + genre_science_fiction + genre_horror +
##     genre_drama + genre_animation + genre_romance + genre_western +
##     genre_history + genre_music
##
##                         Df  Sum of Sq        RSS    AIC
## - genre_music            1 1.3502e+16 4.8200e+19 160419
## - genre_fantasy          1 1.4113e+16 4.8201e+19 160419
## - genre_action           1 1.6769e+16 4.8203e+19 160419
## - genre_war              1 2.1759e+16 4.8208e+19 160419
## <none>                                4.8186e+19 160420
## - genre_crime            1 2.2754e+16 4.8209e+19 160420
## - genre_science_fiction  1 2.5345e+16 4.8212e+19 160420
## - genre_family           1 3.2147e+16 4.8219e+19 160420
## - genre_romance          1 3.4295e+16 4.8221e+19 160421
## - genre_animation        1 3.7559e+16 4.8224e+19 160421
## - genre_thriller         1 4.2695e+16 4.8229e+19 160421
## - genre_history          1 5.3399e+16 4.8240e+19 160422
## - genre_horror           1 6.4407e+16 4.8251e+19 160423
## - genre_documentary      1 8.5914e+16 4.8272e+19 160425
## - genre_western          1 1.0776e+17 4.8294e+19 160427
## - genre_drama            1 1.5037e+17 4.8337e+19 160431
## - genre_adventure        1 1.7907e+17 4.8366e+19 160434
## - average_rating         1 1.4054e+18 4.9592e+19 160542
## - budget                 1 2.5747e+19 7.3934e+19 162276
##
## Step:  AIC=160418.7
## profit ~ budget + average_rating + genre_thriller + genre_war +
##     genre_fantasy + genre_documentary + genre_action + genre_family +
##     genre_crime + genre_adventure + genre_science_fiction + genre_horror +
##     genre_drama + genre_animation + genre_romance + genre_western +
##     genre_history
##
##                         Df  Sum of Sq        RSS    AIC
## - genre_fantasy          1 1.3885e+16 4.8214e+19 160418
## - genre_action           1 1.6911e+16 4.8217e+19 160418
## <none>                                4.8200e+19 160419
## - genre_war              1 2.2521e+16 4.8222e+19 160419
## - genre_crime            1 2.3655e+16 4.8224e+19 160419
## - genre_science_fiction  1 2.6132e+16 4.8226e+19 160419
## - genre_family           1 3.3728e+16 4.8234e+19 160420
## - genre_romance          1 3.6153e+16 4.8236e+19 160420
## - genre_animation        1 3.7868e+16 4.8238e+19 160420
## - genre_thriller         1 4.4933e+16 4.8245e+19 160421
## - genre_history          1 5.4713e+16 4.8255e+19 160422
## - genre_horror           1 6.3557e+16 4.8264e+19 160422
## - genre_documentary      1 8.1386e+16 4.8281e+19 160424
## - genre_western          1 1.0859e+17 4.8309e+19 160426
## - genre_drama            1 1.4773e+17 4.8348e+19 160430
```

```
## - genre_adventure        1 1.7589e+17 4.8376e+19 160433
## - average_rating         1 1.4054e+18 4.9605e+19 160541
## - budget                 1 2.5748e+19 7.3948e+19 162275
##
## Step:  AIC=160418
## profit ~ budget + average_rating + genre_thriller + genre_war +
##     genre_documentary + genre_action + genre_family + genre_crime +
##     genre_adventure + genre_science_fiction + genre_horror +
##     genre_drama + genre_animation + genre_romance + genre_western +
##     genre_history
##
##                          Df  Sum of Sq        RSS    AIC
## - genre_action            1 1.6532e+16 4.8230e+19 160417
## <none>                                 4.8214e+19 160418
## - genre_war               1 2.3252e+16 4.8237e+19 160418
## - genre_crime             1 2.5236e+16 4.8239e+19 160418
## - genre_science_fiction   1 2.5884e+16 4.8240e+19 160418
## - genre_romance           1 3.6916e+16 4.8251e+19 160419
## - genre_animation         1 3.8241e+16 4.8252e+19 160419
## - genre_family            1 4.1603e+16 4.8255e+19 160420
## - genre_thriller          1 4.8050e+16 4.8262e+19 160420
## - genre_history           1 5.6773e+16 4.8271e+19 160421
## - genre_horror            1 6.7406e+16 4.8281e+19 160422
## - genre_documentary       1 8.3849e+16 4.8298e+19 160424
## - genre_western           1 1.1141e+17 4.8325e+19 160426
## - genre_drama             1 1.4992e+17 4.8364e+19 160429
## - genre_adventure         1 1.9023e+17 4.8404e+19 160433
## - average_rating          1 1.4064e+18 4.9620e+19 160541
## - budget                  1 2.6517e+19 7.4730e+19 162318
##
## Step:  AIC=160417.5
## profit ~ budget + average_rating + genre_thriller + genre_war +
##     genre_documentary + genre_family + genre_crime + genre_adventure +
##     genre_science_fiction + genre_horror + genre_drama + genre_animation +
##     genre_romance + genre_western + genre_history
##
##                          Df  Sum of Sq        RSS    AIC
## <none>                                 4.8230e+19 160417
## - genre_war               1 2.5699e+16 4.8256e+19 160418
## - genre_crime             1 3.1760e+16 4.8262e+19 160418
## - genre_science_fiction   1 3.2426e+16 4.8263e+19 160418
## - genre_animation         1 4.0872e+16 4.8271e+19 160419
## - genre_romance           1 4.3366e+16 4.8274e+19 160419
## - genre_family            1 5.2806e+16 4.8283e+19 160420
## - genre_history           1 5.9191e+16 4.8290e+19 160421
## - genre_thriller          1 6.0641e+16 4.8291e+19 160421
## - genre_documentary       1 7.8086e+16 4.8308e+19 160422
## - genre_horror            1 7.9756e+16 4.8310e+19 160423
## - genre_western           1 1.1148e+17 4.8342e+19 160425
## - genre_drama             1 1.3807e+17 4.8368e+19 160428
## - genre_adventure         1 1.7441e+17 4.8405e+19 160431
## - average_rating          1 1.4318e+18 4.9662e+19 160542
## - budget                  1 2.6963e+19 7.5194e+19 162343
```

```
##
## Call:
## lm(formula = profit ~ budget + average_rating + genre_thriller +
##     genre_war + genre_documentary + genre_family + genre_crime +
##     genre_adventure + genre_science_fiction + genre_horror +
##     genre_drama + genre_animation + genre_romance + genre_western +
##     genre_history, data = movies)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -562510331  -33506994   -9116101   19314364 1994305666
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -1.087e+08  1.102e+07  -9.860  < 2e-16 ***
## budget                      2.151e+00  4.375e-02  49.172  < 2e-16 ***
## average_rating              3.898e+07  3.440e+06  11.331  < 2e-16 ***
## genre_thrillerTRUE         -9.897e+06  4.244e+06  -2.332 0.019750 *
## genre_warTRUE              -1.629e+07  1.073e+07  -1.518 0.129074
## genre_documentaryTRUE      -2.307e+07  8.719e+06  -2.646 0.008170 **
## genre_familyTRUE            1.413e+07  6.493e+06   2.176 0.029603 *
## genre_crimeTRUE            -8.586e+06  5.087e+06  -1.688 0.091556 .
## genre_adventureTRUE         2.124e+07  5.370e+06   3.955 7.78e-05 ***
## genre_science_fictionTRUE -9.949e+06  5.834e+06  -1.705 0.088224 .
## genre_horrorTRUE            1.665e+07  6.226e+06   2.674 0.007516 **
## genre_dramaTRUE            -1.342e+07  3.813e+06  -3.519 0.000438 ***
## genre_animationTRUE         1.507e+07  7.874e+06   1.914 0.055626 .
## genre_romanceTRUE           8.693e+06  4.408e+06   1.972 0.048672 *
## genre_westernTRUE          -5.952e+07  1.882e+07  -3.162 0.001579 **
## genre_historyTRUE          -2.154e+07  9.351e+06  -2.304 0.021276 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 105600000 on 4325 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.485,  Adjusted R-squared:  0.4832
## F-statistic: 271.5 on 15 and 4325 DF,  p-value: < 2.2e-16
```

Discussion

Challenges

Not robust to re-sampling. Co-linearity between genres