# Director’s Analysis

## Abstract/Summary

This purpose of this separate analysis is to look at the dataset with a focus on directors who have more than 3 films in their filmography with respect to our dataset. The goal is to find a prediction equation for predicting box office gross with a focus on directors.

Similar to the other models in this report, the response variable of adjusted gross was transformed to log(adjgross).

The best prediction equation, after residual plots, variable selection and using criteria of adjusted R2 and cross-validated root mean square prediction error is:

log(adjgross) = 8.200 – 0.077lncast\_total\_facebook\_likes – 0.0421lnactor1\_facebook\_likes + 0.061lndirector\_facebook\_likes + 1.127lnnum\_voted\_users – 0.358lnnum\_users\_for\_review + 0.250lnbudget – 0.301imdb\_score – 0.675content\_ratingPG-13 – 0.791content\_ratingR + 0.004duration +0.59BobbyFarrelly + 0.996BrettRatner + 0.794ClintEastwood – 0.098DavidFincher + 0.915JoelSchumacher + 0.794JohnMcTiernan – 0.137MartinScorsese + 0.151PaulWSAnderson + 0.564RennyHarlin + 1.084RichardDonner + 1.41RobCohen + 0.53RonHoward + 0.789ShawnLevy + 0.775StephenFrears + 0.884StevenSpielberg + 0.475TimBurton + 0.464TonyScott

Note that director categorical variables are not grouped together. This is because directors act like a brand, and grouping them together would not provide any useful analysis as directors do not work together in real life.

Significant categorical variables include Steven Spielberg, Rob Cohen, Clint Eastwood and Brett Ratner. A review of their films show that their films are not only typically high grossing but very profitable.

Steven Spielberg adds an average of 0.884 to the log(adjgross), Rob Cohen adds 1.41, Clint Eastwood adds 0.795 and Brett Ratner adds 0.996, holding the baseline constant with Bob Farrelly.

## Description of Data

Data are collected by scraping both imdb resources and Box Office Mojo websites. Assumptions were made regarding the adjusted gross (estimated data from Box Office Mojo) which are outlined in APPENDIX X.

|  |  |
| --- | --- |
| Variables | Explanation or unit |
| adjgross | Gross US domestic box office, adjusted for inflation |
| director\_facebook\_likes | Number of Facebook likes a director has |
| actor\_1\_facebook\_likes | Number of Facebook likes from the lead actor |
| duration | Duration of the film (in minutes) |
| num\_voted\_users | Number of users who have voted for film on imdb |
| cast\_total\_facebook\_likes | Number of Facebook likes received by the cast |
| Budget | Budget for the film |
| num\_users\_for\_review | Number of uses that have reviewed the film |
| movie\_facebook\_likes  content\_rating  director\_name | Number of Facebook likes received by the film  Rating for the film (e.g. PG, R, NC-17, etc.)  Film directors that have made more than 3 films |

This particular data set subsets the original data set such that only directors that have made more than 3 films are included. This reduces the dataset from over 5000 observations to approximately 400.

Directors are not combined into fewer categorical variables because in practical situations, directors do not work together and their contributions to the movie industry are considered singular.

## Data Analysis and Results

Summary statistics are provided are given in Figure X. Plots that justify the use of transforms for explanatory variables are given in Figure Y and remain unchanged through this subset analysis.

SAMPLE CORRELATIONS

We fit a multiple regression model (Table X) with the explanatory variables given above. The adj-R2 of this model is 0.7412 with a Residual Standard Error of 0.6683.

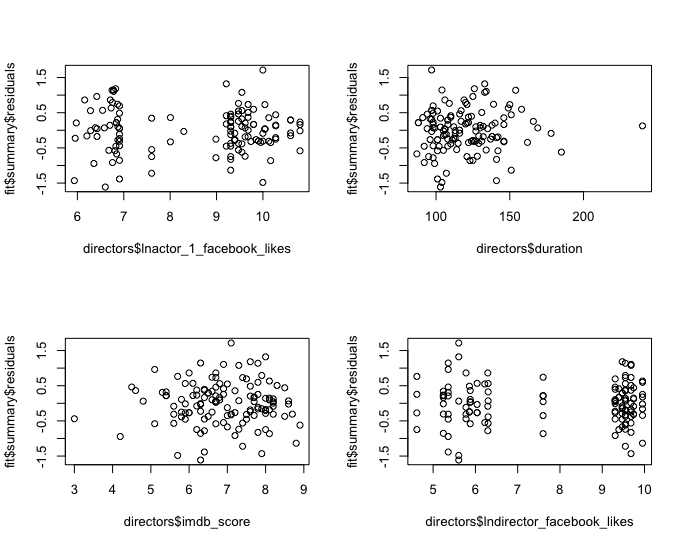
### Residual Plots

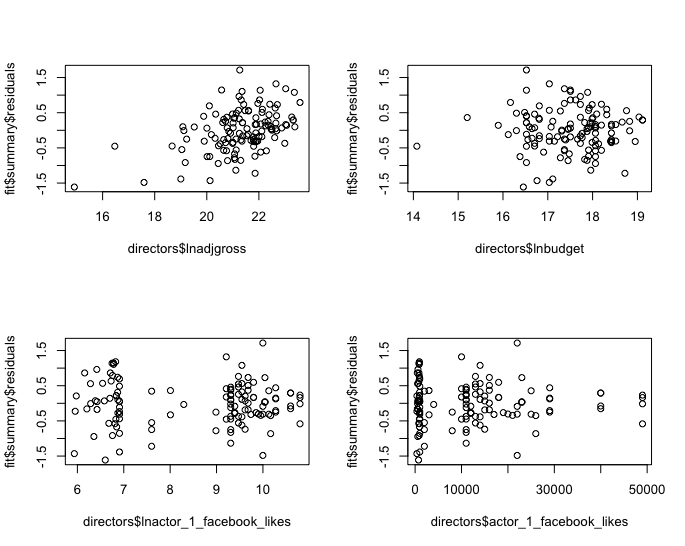
Sample residual plots are given in Figure Z. The residuals do not suggest any curvilinear form, therefore quadratic terms or interaction terms were not added to the model.

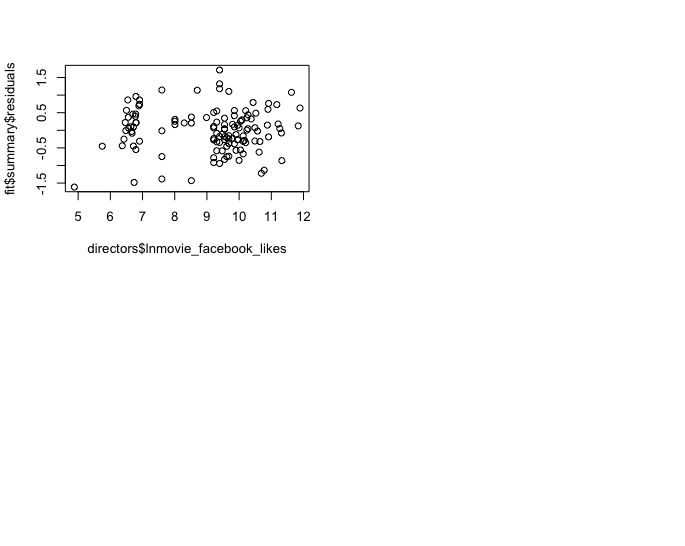
Backward Elimination and Forward Selection techniques were used to determine other good models. The 2nd best model eliminated two explanatory variables: lntotal\_facebook\_likes and lndirector\_facebook\_likes. This model, as well as the full model, are compared via Cp statistic, adj-R2 and error from a random holdout set (Table V).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| variable | Estimate | Std.Error | tvalue | Pr(>|t|) |  |
| (Intercept) | 8.199598 | 2.09028 | 3.923 | 0.000161 | \*\*\* |
| lncast\_total\_facebook\_likes | -0.077328 | 0.202517 | -0.382 | 0.703395 |  |
| lnactor\_1\_facebook\_likes | -0.042137 | 0.160888 | -0.262 | 0.793934 |  |
| lndirector\_facebook\_likes | 0.060706 | 0.100062 | 0.607 | 0.545437 |  |
| lnnum\_voted\_users | 1.127142 | 0.158136 | 7.128 | 1.62E-10 | \*\*\* |
| lnnum\_users\_for\_review | -0.357905 | 0.198152 | -1.806 | 0.073892 | . |
| lnbudget | 0.250056 | 0.103054 | 2.426 | 0.017038 | \* |
| imdb\_score | -0.301169 | 0.119859 | -2.513 | 0.01358 | \* |
| content\_ratingPG-13 | -0.674736 | 0.207765 | -3.248 | 0.001585 | \*\* |
| content\_ratingR | -0.791106 | 0.238341 | -3.319 | 0.00126 | \*\* |
| duration | 0.003759 | 0.004505 | 0.834 | 0.406121 |  |
| director\_nameBobbyFarrelly | 0.590241 | 0.527586 | 1.119 | 0.265923 |  |
| director\_nameBrettRatner | 0.996214 | 0.454824 | 2.19 | 0.030824 | \* |
| director\_nameClintEastwood | 0.794618 | 0.31924 | 2.489 | 0.014457 | \* |
| director\_nameDavidFincher | -0.098208 | 0.405482 | -0.242 | 0.80912 |  |
| director\_nameJoelSchumacher | 0.914539 | 0.472327 | 1.936 | 0.055661 | . |
| director\_nameJohnMcTiernan | 0.793879 | 0.455977 | 1.741 | 0.08475 | . |
| director\_nameMartinScorsese | -0.136908 | 0.460158 | -0.298 | 0.766683 |  |
| director\_namePaulW.S.Anderson | 0.151354 | 0.416308 | 0.364 | 0.716951 |  |
| director\_nameRennyHarlin | 0.56406 | 0.4139 | 1.363 | 0.176009 |  |
| director\_nameRichardDonner | 1.084072 | 0.556756 | 1.947 | 0.054325 | . |
| director\_nameRobCohen | 1.405827 | 0.43309 | 3.246 | 0.001593 | \*\* |
| director\_nameRonHoward | 0.527189 | 0.362782 | 1.453 | 0.149303 |  |
| director\_nameShawnLevy | 0.788692 | 0.451607 | 1.746 | 0.08381 | . |
| director\_nameStephenFrears | 0.774994 | 0.48599 | 1.595 | 0.113942 |  |
| director\_nameStevenSpielberg | 0.883613 | 0.325812 | 2.712 | 0.007874 | \*\* |
| director\_nameTimBurton | 0.474526 | 0.332979 | 1.425 | 0.157244 |  |
| director\_nameTonyScott | 0.463898 | 0.408975 | 1.134 | 0.259383 |  |
|  |  |  |  |  |  |
| Residual standard error: 0.6683 on 100 degrees of freedom | | |  |  |  |
| Multiple R-squared: 0.7962, Adjusted R-squared: 0.7412 | | |  |  |  |
| F-statistic: 14.47 on 27 and 100 DF, p-value: < 2.2e-16 | | |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Univariate Summary Statistics** | | | |  | |  | |  | |  | |
|  | lnnum\_voted\_users | | lncast\_total\_facebook\_likes | | lnbudget | | lnnum\_users\_for\_review | | lnmovie\_facebook\_likes | |
| Min. | 7.243 | | 6.252 | | 14.08 | | 3.258 | | 4.883 | |
| 1st Q | 10.587 | | 8.12 | | 16.81 | | 5.13 | | 7.905 | |
| Median | 11.789 | | 9.483 | | 17.5 | | 5.807 | | 9.547 | |
| Mean | 11.644 | | 9.177 | | 17.5 | | 5.826 | | 9.079 | |
| 3rd Q | 12.615 | | 10.096 | | 18.06 | | 6.539 | | 10.127 | |
| Max. | 14.114 | | 11.3 | | 19.11 | | 7.996 | | 11.891 | |
|  |  | |  | |  | |  | |  | |
|  | lnadjgross | | lngross | | lndirector\_facebook\_likes | | lnactor\_1\_facebook\_likes | | duration | |
| Min. | 14.92 | | 11.23 | | 4.615 | | 5.935 | | 87 | |
| 1st Q | 20.77 | | 17.08 | | 5.858 | | 6.872 | | 103 | |
| Median | 21.38 | | 17.69 | | 9.306 | | 9.306 | | 116 | |
| Mean | 21.33 | | 17.64 | | 7.858 | | 8.575 | | 119.8 | |
| 3rd Q | 22.12 | | 18.44 | | 9.547 | | 9.695 | | 132 | |
| Max. | 23.58 | | 19.89 | | 9.952 | | 10.8 | | 240 | |
|  |  | |  | |  | |  | |  | |
| **Frequency Table for Directions with >= 3 Movies** | | | | | |  | |  | |  | |
| Barry Levinson | | Bobby Farrelly | | Brett Ratner | | Clint Eastwood | | David Fincher | |  | |
| 5 | | 4 | | 4 | | 13 | | 8 | |  | |
| Joel Schumacher | | John McTiernan | | Martin Scorsese | | Paul WS Anderson | | Renny Harlin | |  | |
| 3 | | 4 | | 6 | | 5 | | 9 | |  | |
| Rob Cohen | | Ron Howward | | Shaw Levy | | Stephen Frears | | Steven Spielberg | |  | |
| 6 | | 6 | | 8 | | 3 | | 17 | |  | |
| Tim Burton | | Tony Scott | | Woody Allen | |  | |  | |  | |
| 9 | | 4 | | 12 | |  | |  | |  | |
|  | |  | |  | |  | |  | |  | |
| **Frequency Table for Content Rating** | | | | | |  | |  | |  | |
| Approved | | G | | GP | | M | | NC-17 | |  | |
| 303 | | 55 | | 112 | | 6 | | 5 | |  | |
| PG-13 | | R | | TV-14 | | TV-G | | TV-MA | |  | |
| 1461 | | 2118 | | 30 | | 10 | | 20 | |  | |
|  | |  | |  | |  | |  | |  | |
| Not | | Rated | | Passed | | PG | | X | |  | |
| 7 | | 116 | | 9 | | 701 | | 13 | |  | |
| TV-PG | | TV-Y | | TV-Y7 | | Unrated | |  | |  | |
| 13 | | 1 | | 1 | | 62 | |  | |  | |







|  |  |  |
| --- | --- | --- |
| statistic / model | fit1 | fit2 |
| Adj-R2 | 0.7962 | 0.7945 |
| Res-SD | 0.6683 | 0.6646 |
| rmsepred(leave-one-out) | 0.834 | 0.82 |
| rmsepred(train / holdout) | 1.887558 | ?? |

## Brief Discussion:

In conclusion, we have found a good-fitting model and residual plots that show no sign of curvilinear form. The sample did not include foreign films – only domestic films made in the US were used for this subset.

It is possible to include genre data or actor data, but the model may suffer from overfitting if the categorical data is not grouped into fewer groups.

The RMSE for the training/holdout is quite high. It’s possible that there could be some issues with the data, or that simply it is very hard to predict the performance of films, and high variability is to be expected when doing an analysis within this industry.