# Actor’s Analysis

## Abstract/Summary

This purpose of this separate analysis is to look at the dataset with a focus on actors who have performed in more than 25 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 actors.

The response variable of adjusted gross was transformed to log(adjgross), otherwise the residual plots seems not heteroscedastic .

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) = -2.0176274- -0.0062244duration

+ 0.8828473lognum\_voted\_users -0.0031701num\_critic\_for\_reviews

+ 0.5978969logbudget + 0.1475909 lognum\_user\_for\_reviews+ 0.0233070 imdb\_score + 0.6570164 Denzel Washington + 0.7746052Harrison Ford+ 0.4390425 J.K. Simmons -0.5567939 Jason Statham-0.0978200 Johnny Depp +0.0513779 Liam Neeson + 0.1420526 Matt Damon -0.0958409 Nicolas Cage -0.0127797 Robert De Niro + 0.1114163 Robert Downey Jr.+ 0.3149969 Robin Williams -0.1972292 content\_ratingPG -0.4430192 content\_ratingPG-13 -0.8910856 content\_ratingR+ 1.3003917 content\_ratingX

Actor categorical variables are not grouped together. Since an actor him/herself is a brand, and we only focus on the effect of a single main actor on the gross , so grouping them together would not provide any useful analysis.

Significant categorical variables of actors include Denzel Washington

, Harrison Ford and Jason Statham.

Denzel Washington adds an average of 0.657to the log(adjgross), Harrison Ford adds 0.775, and Jason Statham minus 0.5567939, holding the baseline constant with Bruce Willis

## 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 |
| 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\_critic\_for\_review | Number of uses that have reviewed the film |
| movie\_facebook\_likes  content\_rating  actor\_name | Number of Facebook likes received by the film  Rating for the film (e.g. PG, R, NC-17, etc.)  Actors who has joined more than 25 films |

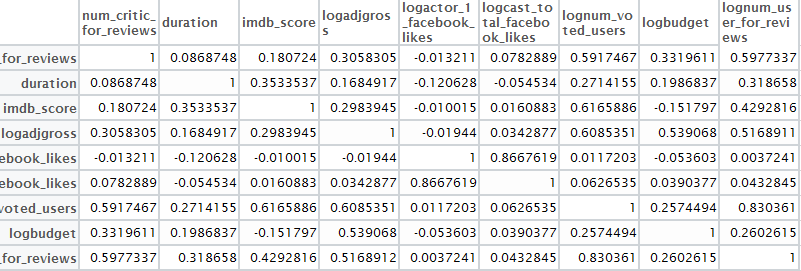
This particular data set subsets the original data set such that only actors that have joined more than 25 films. This reduces the dataset from over 5000 observations to 356.

Actors are not combined into fewer categorical variables because we try to focus on the effect of a single actor so their contributions to the gross 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 best fit model suggested by both adj and cp value eliminated two explanatory variables: logactor\_1\_facebook\_likes and logcast\_total\_facebook\_likes. This model, as well as the full model, are compared via Cp statistic, adj-R2 Cross-validation root mean square error(leave-one-out) and Cross-validation root mean square error from a random holdout set (Table V).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| lm(formula = logadjgross ~ duration + lognum\_voted\_users + num\_critic\_for\_reviews + | | | | | | | | | | |
| logbudget + lognum\_user\_for\_reviews + imdb\_score + actor\_1\_name + | | | | | | | | |  |  |
| content\_rating, data = ac1) | | | |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| Residuals: | |  |  |  |  |  |  |  |  |  |
| Min 1Q Median 3Q Max | | | | | |  |  |  |  |  |
| -6.0444 -0.3224 0.0450 0.4886 3.5413 | | | | | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| Coefficients: | |  |  |  |  |  |  |  |  |  |
| variable | Estimate | Std. Error | t value | Pr(>|t|) |  |  |  |  |  |  |
| (Intercept) | -2.01763 | 1.453236 | -1.388 | 0.16595 |  |  |  |  |  |  |
| duration | -0.00622 | 0.002684 | -2.319 | 0.02097 | \* |  |  |  |  |  |
| lognum\_voted\_users | 0.882847 | 0.123172 | 7.168 | 4.93E-12 | \*\*\* |  |  |  |  |  |
| num\_critic\_for\_reviews | -0.00317 | 0.000682 | -4.648 | 4.84E-06 | \*\*\* |  |  |  |  |  |
| logbudget | 0.597897 | 0.062404 | 9.581 | < 2e-16 | \*\*\* |  |  |  |  |  |
| lognum\_user\_for\_reviews | 0.147591 | 0.134108 | 1.101 | 0.27189 |  |  |  |  |  |  |
| imdb\_score | 0.023307 | 0.096751 | 0.241 | 0.80978 |  |  |  |  |  |  |
| actor\_1\_nameDenzel Washington | 0.657016 | 0.267544 | 2.456 | 0.01457 | \* |  |  |  |  |  |
| actor\_1\_nameHarrison Ford | 0.774605 | 0.280838 | 2.758 | 0.00613 | \*\* |  |  |  |  |  |
| actor\_1\_nameJ.K. Simmons | 0.439043 | 0.265523 | 1.654 | 0.09917 | . |  |  |  |  |  |
| actor\_1\_nameJason Statham | -0.55679 | 0.281165 | -1.98 | 0.04849 | \* |  |  |  |  |  |
| actor\_1\_nameJohnny Depp | -0.09782 | 0.254302 | -0.385 | 0.70073 |  |  |  |  |  |  |
| actor\_1\_nameLiam Neeson | 0.051378 | 0.276059 | 0.186 | 0.85247 |  |  |  |  |  |  |
| actor\_1\_nameMatt Damon | 0.142053 | 0.279337 | 0.509 | 0.61141 |  |  |  |  |  |  |
| actor\_1\_nameNicolas Cage | -0.09584 | 0.263541 | -0.364 | 0.71634 |  |  |  |  |  |  |
| actor\_1\_nameRobert De Niro | -0.01278 | 0.24486 | -0.052 | 0.95841 |  |  |  |  |  |  |
| actor\_1\_nameRobert Downey Jr. | 0.111416 | 0.282508 | 0.394 | 0.69355 |  |  |  |  |  |  |
| actor\_1\_nameRobin Williams | 0.314997 | 0.283612 | 1.111 | 0.26751 |  |  |  |  |  |  |
| content\_ratingPG | -0.19723 | 0.739032 | -0.267 | 0.78973 |  |  |  |  |  |  |
| content\_ratingPG-13 | -0.44302 | 0.736631 | -0.601 | 0.54797 |  |  |  |  |  |  |
| content\_ratingR | -0.89109 | 0.737744 | -1.208 | 0.22796 |  |  |  |  |  |  |
| content\_ratingX | 1.300392 | 1.053237 | 1.235 | 0.21782 |  |  |  |  |  |  |
| --- |  |  |  |  |  |  |  |  |  |  |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 | | | | | | | |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| Residual standard error: 0.9982 on 334 degrees of freedom | | | | | | | |  |  |  |
| Multiple R-squared: 0.644, Adjusted R-squared: 0.6216 | | | | | | | |  |  |  |
| F-statistic: 28.77 on 21 and 334 DF, p-value: < 2.2e-16 | | | | | | | |  |  |  |

**Univariate Summary Statistics**

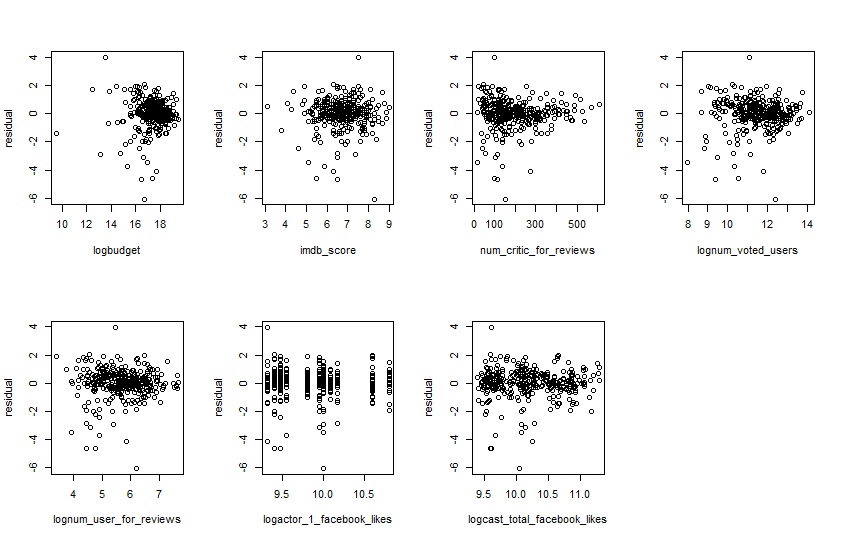
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| num\_critic\_for\_reviews | duration | imdb\_score | logadjgross | logactor\_1\_facebook\_likes | |
| Min. : 16.0 | Min. : 77.0 | Min. :3.100 | Min. : 9.709 | Min. : 9.306 |  |
| 1st Qu.:112.0 | 1st Qu.:103.0 | 1st Qu.:6.200 | 1st Qu.:17.184 | 1st Qu.: 9.473 |  |
| Median :167.0 | Median :113.5 | Median :6.700 | Median :17.956 | Median : 9.952 |  |
| Mean :196.6 | Mean :117.3 | Mean :6.722 | Mean :17.686 | Mean : 9.903 |  |
| 3rd Qu.:266.2 | 3rd Qu.:126.0 | 3rd Qu.:7.300 | 3rd Qu.:18.594 | 3rd Qu.:10.106 |  |
| Max. :608.0 | Max. :289.0 | Max. :9.000 | Max. :21.047 | Max. :10.800 |  |
| lognum\_user\_for\_review | s | logcast\_total\_facebook\_likes | lognum\_voted\_users | logbudget |  |
| Min. :3.401 | | Min. : 9.384 | Min. : 7.979 | Min. : 9.547 |  |
| 1st Qu.:5.112 | | 1st Qu.: 9.726 | 1st Qu.:10.770 | 1st Qu.:16.860 |  |
| Median :5.768 | | Median :10.136 | Median :11.562 | Median :17.504 |  |
| Mean :5.710 | | Mean :10.188 | Mean :11.503 | Mean :17.390 |  |
| 3rd Qu.:6.260 | | 3rd Qu.:10.596 | 3rd Qu.:12.299 | 3rd Qu.:18.081 |  |
| Max. :7.694 | | Max. :11.300 | Max. :14.097 | Max. :19.519 |  |

**Frequency Table for actors involved >= 25 Movies**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | |  |  |  |
| Bruce Willis | Denzel Washington | Harrison Ford | J.K. Simmons | Jason Statham |  |
| 29 | 30 | 25 | 31 | 25 |  |
| Johnny Depp | Liam Neeson | Matt Damon | Nicolas Cage | Robert De Niro |  |
| 39 | 26 | 28 | 30 | 42 |  |
| Robert Downey Jr. | Robin Williams |  |  |  |  |
| 26 | 25 |  |  |  |  |
|  |  |  |  |  |  |

**Frequency Table for Content Rating**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| PG-13 | G | PG | R | X |  |
| 127 | 2 | 44 | 181 | 2 |  |
|  |  |  |  |  |  |



|  |
| --- |
| full model fit by cp |
| adjusted R2 0.5557362 0.5580501 |
| residual SD 1.0815552 1.0787350 |
| rmsepred(leave-one-out) 1.1029022 1.0972481 |
| rmsepred(train/holdout) 1 1.0892882 1.0513796 |
| rmsepred(train/holdout) 2 1.0817213 1.2142132 |
| rmsepred(train/holdout) 3 1.0732342 1.0552653 |

## Brief Discussion:

The model we found for actors is fine, there is no sign of curvilinear form.

We treat the data separately for actor and director, since too many category variables may cause overfitting and looks messy. The adjusted R2 for the best model we choose is about 0.56, which is not very high, however the RMSE for train/holdout is really high , indicating that it may be hard to predict the gross of films using the data given and high variability is to be expected when doing an analysis within this industry. One thing to be noticed, when fitting the best model selected by regsubsets, most of the variables we kept are not related to actors and only 3 of the actors shows statistical significance to the response variable, which seems a little bit unreasonable.