Influential Factors for Gross Profit in Film

Hugh Chen, Tianqi Zhang, Yao Xiao

University of Washington

Research Question

Controlling for certain factors, can we meaningfully interpret the effect of estimators of reputation (scores) and theme (genres) on gross profit?

Data

Objective: Description and Inference

Data Collection

- IMDB dataset from Kaggle - IMDB data from a Kaggle dataset which scraped 5000+ movies with 28 movie-related variables from IMDB website.
- Rotten Tomatoes data by scraping - Rotten Tomatoes scores by "tomatometer" score from the review pages for the movies.

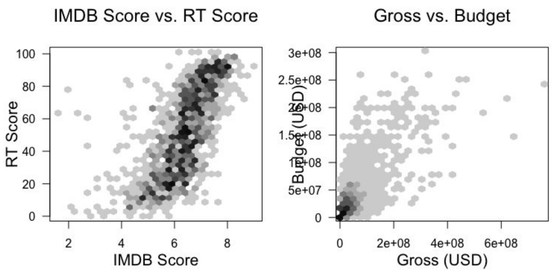
Data Set

We selected movies among the year 2006 to 2015 and made in the US to avoid possible currency differences due to country. In total we have 1329 data points.

- Control Variable: Year - to control for possible economic factors
- Dependent Variable: Gross Profit (USD)
- Independent Variables
  - Budget (USD)
  - IMDB scores- Range from 1 to 10
  - Rotten Tomatoes scores - Range from 0 to 100%
  - Genre: 21 levels in total
    Action Adventure Animation Biography Comedy Crime Documentary Drama Family Fantasy History Horror Music Musical Mystery Romance Sci-Fi Sport Thriller War Western

Exploratory Analysis

Non-linear relationships in bivariate scatter plots (using hexagonal binning).



Transformations

On the left we have a logistic shaped curve and on the right we have long tails. In order to combat this, we applied multivariate Box-Cox transformations to get the following λs:

Variable	Gross	Budget	IMDB Score	RT Score
λ	0.295	0.255	2.403	0.791

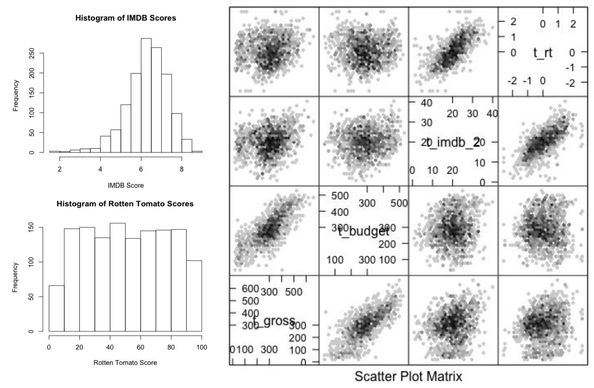
Data Cont'd

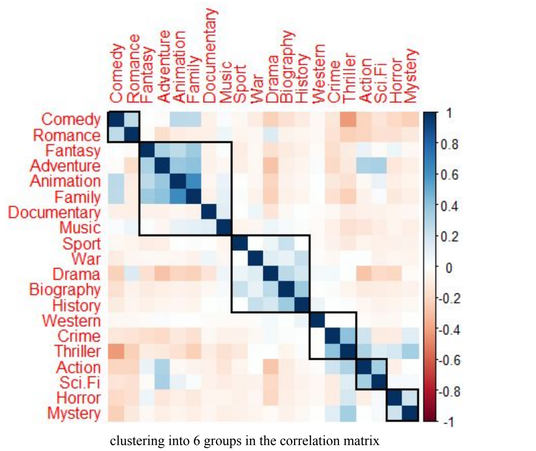
Transformations Cont'd

Upon investigating our transformed scatter plots, we see ellipsoids that suggest linearity between most variables except for with Rotten Tomatoes scores. Since we see a relatively uniform distribution for the Rotten Tomato scores, we decided to transform the data using a normal inverse CDF, to achieve linearity. We use the following transformations (achieved by finding the multivariate Box-Cox transformation after transforming rotten tomatoes scores):

Variable	Gross	Budget	IMDB Score	RT Score
Transform	λ=0.25	λ=0.25	λ=2	qnorm

Then, our resulting pairwise scatterplots were fairly linear:





clustering into 6 groups in the correlation matrix

Results

Model Selection

Our genre factor has 21 levels that are not exclusive. Based on the initial regression result (see handout), most of those variables are not significant. We want to reduce the numbers of genres and derive new variables from them for better interpretation.

First we combined Music and Musical as one variable - it's just two sayings for the same theme. Then we used hierarchical clustering and tried to separate the 20 variables into different numbers of groups. We prefer simpler model, so BIC Criterion is used to check which one is better:

3 groups	4 groups	5 groups	6 groups	7 groups	Raw Data
15323.41	15320.90	15323.15	15314.63	15321.64	15361.32

So we choose to reduce genre into six binary variables as indicated in the correlation matrix. The six groups from hierarchical cluster also seems reasonable by intuition:

Comedy|Romance: Date Night
Fantasy|Adventure|Animation|Family|Docu.|Music: Family Oriented
Sport|War|Drama|Biography|History: Heavy Subjects
Western|Crime|Thriller: Old Fashioned Excitement
Action|Sci.Fi: Science Fiction
Horror|Mystery: For Curious Minds

Regression Result

$Gross^{0.25} \sim$ as.factor(Year) + $Budget^{0.25}$ + IMDB + $IMDB^2$ + qnorm(RT) + c1+c2+c3+c4+c5+c6

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	186.21	42.27	4.40	<0.0001	***
Year	(11 factors)				
$Budget^{0.25}$	0.99	0.03	37.56	<0.0001	***
$IMDB^2$	6.94	1.22	5.70	<0.0001	***
IMDB	-73.99	14.15	-5.23	<0.0001	***
qnorm(RT)	13.94	3.48	4.00	<0.0001	***
c1	16.52	4.90	3.37	0.0008	***
c2	-4.69	4.76	-0.98	0.33	
c3	-20.49	6.75	-3.04	0.0024	**
c4	-4.31	5.18	-0.83	0.41	
c5	-6.84	5.43	-1.26	0.21	
c6	25.46	6.10	4.17	<0.0001	***

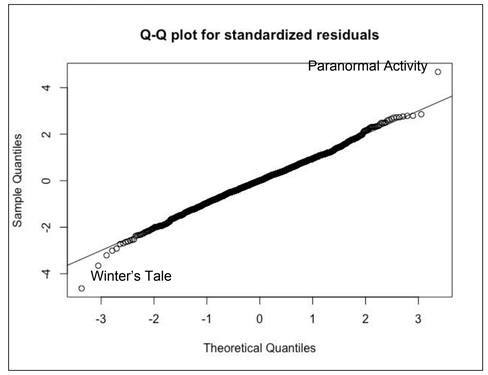
Residual standard error: 73.22 on 1309 degrees of freedom
Multiple R-squared: 0.6232, Adjusted R-squared: 0.6177
F-statistic: 113.9.7 on 19 and 1309 DF, p-value: < 2.2e-16

Censor:

Our residuals and univariate summaries for our transformed variables suggested that we might have a natural censor for our dependent variable. We fit a tobit regression model to see if this censor made any significant difference and we found that the coefficients of interest in our model didn't change much at all (see handout).

Diagnostics

We identified 2 outliers:



Title	Genres	Gross ($)	Budget ($)	Year	IMDB Score	RT Score	Std.res
Winter's Tale	Drama|Fantasy|Mystery|Romance	22,451	60,000,000	2014	6.2	13	-4.63
Paranormal Activity	Horror	107,917,283	15,000	2007	6.3	83	4.67

Cook's distances of our model are all below 0.025.
Deletion of cases with large Cook's distance didn't change the coefficient estimates by much(see handout).

Conclusions

- Both IMDB score and Rotten Tomatoes score are significantly correlated with movie gross after controlling for other variables. The effect of IMDB score has a quadratic form. Since IMDB scores are from general public and tomatometer are mostly from professional film critics, these two scores may share something in common but still carry different information.
- Genre groups that are Comedy|Romance or Horror|Mystery are significantly positively associated with gross profit while movies with Sport|War|Drama|Biography|History genre tend to have less effect on gross profit conditional on other variables. This might suggest that it's more profitable for filmmakers to focus on romantic comedies or scary mystery movies.

References

- https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset
- https://www.rottentomatoes.com/
- https://www.imdb.com/