EDU S040: Introductory and Intermediate Statistics for Educational Research

Final Project

**Assessing the relationship between movie revenues and news coverage**

1. **Introduction**

Considering only the market in the United States, Cinema is a multi-billion dollar industry. Film producers and directors pour millions and months of investment into crafting an entertainment experience with the hopes that it will gain the liking of an audience. If they are successful, they will make a lot of money, but that does not always happen - movies with very little investment ($15,000) such as Paranormal Activity generated hundreds of millions in revenues ($108,000,000), whereas movies such as Battleship with tens of millions in investment ($209,000,000), incurred massive losses (grossing only $65,200,000). These two examples are not uncommon in the film industry.

Given that substantial investments do not guarantee a positive response from the public, it is essential for production houses to understand which other factors contribute to the financial success of a movie. If there is a way to narrow the number of variables that are more closely associated with the success of a film in the US market, this will facilitate investment decision making for executives to reduce the risk of failure and increase the likelihood of producing profitable movies.

Considering that advertising campaigns contribute heavily to inflate the total budget of a movie (Nithin, Pranav, Sarath Babu, and Lijiya, 2014), decisions about how to spend the money destined to advertisements are of paramount importance. Today, news articles about movies constitute an important source of exposure to the public that can be cheaper than other forms of advertising. Moreover, news articles also have the advantage that they are not necessarily perceived as an explicit attempt to persuade the reader. In fact, they often offer negative opinions about a movie. However, it is possible that even negative articles are associated with a positive effect on consumer behavior.

Therefore, understanding the relationship between news articles published and the commercial success of a movie could be of great interest for film producers.  Large data sets like IMDb allow the possibility of investigating such relationships through statistical modeling. In the present study, we are interested in exploring the following research question:

*Adjusting for budget, awards won, and genre popularity what is the association between the number of news articles published about a movie and how much money that movie grosses?*

It is important to consider that the data available about news articles published does not discriminate between positive and negative content about the movie. This fact is not a problem in our research given that we part from the assumption that news articles could have a positive financial impact for the film regardless of their content. In line with this idea, we have developed the following hypothesis:

*There exists a positive association between the number of news articles published about a movie and how much that movie gross in USD, while holding constant for budget, number of awards received, and genre being action, adventure or both.*

We begin by describing the database from which the sample was obtained as well as the variables and techniques we use to conduct the analysis. Then, we present the results and interpretation of our models. Finally, we discuss the implications of the findings and acknowledge the limitations of the study.

**2. Method**

***2.1 Sample***

We started by downloading, cleaning and combining two datasets covering over 15,000 unique titles on IMDb. The first dataset holding 14,762 movie titles was downloaded from Kaggle.com[[1]](#footnote-1); where the purpose of its curation was to create a movie recommendation app. The second dataset containing 5,043 titles was downloaded from data.world[[2]](#footnote-2); where it was aggregated to contain details on traditional and social media coverage of these movie titles.

After cleaning the data and merging the files by unique movie name, we ended up with 909 titles that have no missing values. We then decided to include only titles from after the year 2005 to avoid dealing with large changes in consumer price index, and we also removed all non-US titles and those not presented in USD currency. This left us with our final dataset of 339 observations.

***2.2 Measures***

To answer the research question and test our predictions, we used the following variables:

* **l2gross:** *“Movie gross”* is the outcome variable. This continuous, log-transformed variable indicates how much money a movie made in US dollars.
* **l2news:** *“News articles”* is our key predictor. This continuous log-transformed variable represents the number of articles published about a movie. As mentioned above, this variable does not differentiate between articles with positive and negative content about the movie.
* **l2budget:** *“Movie budget”* is our key control variable. This continuous log-transformed variable indicates the movie’s total production budget.
* **l2wins:** *“Number of awards”* is another control variable. This continuous log-transformed variable represents the number of awards a movie received.
* **act\_adv:** *“Action”* and *“Adventure”* genres are categorical control variables; combined into one *“act\_adv”*. This binary variable indicates if a movie belongs to either genre of action or adventure or both (1) or not (0).

Given that the distribution of the outcome and two of the continuous predictors (*Movie gross*, *Movie budget*, and *Number of awards*) were highly skewed and contained very large ranges, we decided to log-transform (log2) the three variables to make them more interpretable and to meet the inferential assumptions.

***2.3 Analysis***

STATA/IC 15.1 was used to clean and analyze the datasets. We follow a hypothesis driven strategy to initiate our model building. Our plan was to begin with a very basic model, then introduce interactions and/or variable modifications (e.g. squaring, log transforming, etc.) that we thought might improve our model. We would then examine the results (e.g. taxonomy table, correlations, etc.) and conduct refinements as we saw fit. Some of these decisions were done subjectively.

To measure the association between news article coverage and how much a movie grosses, we conducted a multi-linear regression on our final sample set of 339 movies. The formula regressed l2gross (outcome variable) on l2news (key predictor), l2budget (control 1), l2wins (control 2), and act\_adv (binary control).

This regression formula - with its main predictor and controls - was arrived to after several iterations and consultation with S040 teaching staff. We attempted different combinations of variables and interactions that leverage the information in our dataset. Other models tried were eliminated for various reasons; the most prominent of which is that we felt they did not address our research question well enough compared to our chosen model. We also concluded that the log transformations were necessary owing to the shape of the distribution of the variables and their large range.

We decided to control for the budget variable, l2budget, since higher budget movies are reasonably expected to bring in more money to its makers. Our second control, l2wins, was chosen based on the impact awards won may have on moviegoers deciding to see the movie, and thus contributing to its gross. We also selected to control for action and adventure movies for 3 reasons: first, in our exploratory analysis, these genres outperformed all other genres on how much they gross per movie on average; second, action and adventure genres have significantly higher correlations with l2gross compared to all other genre correlations (this point is closely linked to the first point); third, those two genres comprise 123 observations (~36.3%) of our dataset.

After settling on these controls, we tested whether there was an association between how much a movie grosses and the number of news articles written about it.

**3. Results**

***3.1 Univariate statistics***

Table 1 in the Appendix displays the univariate descriptive statistics of all the variables used in this research including log-transformed variables.

The distribution of the outcome variable (movie gross) is unimodal, asymmetrical, and it is extremely skewed to the right (2.17). It has a mean of $63,600,000 and a standard deviation of $77,000,000. Seventy-five percent of the films in the sample had grossed less than $90,000,000 dollars while the top 1% have grossed more than $319,000,000. These outliers are pulling the distribution to the right, which indicates the need to log-transform the variable for better interpretability.

Similarly, the key predictor (news articles) has an asymmetrical unimodal distribution that is extremely skewed to the right (4.02). It has a mean of 1957.51 and a standard deviation of 3054.99. Seventy-five percent of the movies in the sample have less than 2338 news articles published while the top 1% have more than 17,229. For this reason, we also decided to log-transform this variable.

The key control variable (movie budget) has a unimodal distribution that is extremely skewed to the right (1.59) with a mean of $45,100,000 and a standard deviation of $49,600,000. Ninety percent of the movies in the sample have a budget of less than $125,000,000 while the top 1% have a budget greater than $209,000,000. This explains the extreme skewness of the distribution, for which we considered to log-transform the variable.

The additional control variable, (number of awards) has an asymmetrical unimodal distribution that is extremely skewed to the right (3.04). It has a mean of 9.40 and a standard deviation of 18.42. Seventy-five percent of the movies in the sample have won less than 10 awards while the top 1% have won more than 94 awards. As with the other predictors and the outcome, we decided to log-transformed the variable better interpretability.

***3.2 Bivariate statistics***

After analyzing the distribution of the variables, we created a correlation matrix to examine the bivariate relationship between the outcome and all the predictors (see Table 2 in the Appendix). The outcome (logged movie gross) and the key predictor (logged news articles) were highly positively correlated (0.55). This provided initial evidence about the association between the number of published articles and the amount of money a movie grossed. Not surprisingly, the key control variable (logged movie budget) was also highly correlated with the outcome (0.61). Another important observation is that there was a moderate positive linear association between logged movie budget and logged news published (0.33), which indicates that budget by itself cannot account for the number of news published about a movie. Finally, the number of awards was positively correlated with logged gross (0.23) and with logged budget (0.41). All the correlations mentioned above were statistically significant (p < .05).

***3.3 Model building process***

***3.3.1 fitted models***

We began our model selection process by conducting a simple linear regression of movie gross *“gross”* over number of news articles *“news\_articles”* written about it. The result of this regression shows a statistically significant association between the two variables (p<.001). We then considered the shape of the distributions of both *“gross”* and *“news\_articles”*, and decided that log transforming both variables would be apposite as explained in an earlier section of this paper. We then repeated the regression on the log transformed variables, and the result showed that the model was still statistically significant (p<.001).

At this point, we started incorporating controls (including interaction, binary, and continuous variables) that were relevant to our research questions. We began by including an interaction term that reflects if a movie’s genre is action, adventure or both. We also added the log2 transformation of a movie’s budget. Then, we improved on the model by adding the log2 transformation of number of awards won. We were satisfied by the substantive results of this model, and ultimately decided to use it; however, we experimented with a few more models before finally settling on it. (Table.3) in the appendix shows the taxonomy of model fitting.

Other models not selected included introducing an interaction of a movie’s IMDb rating multiplied by the rating count (i.e. how many people rated the movie). We then log2 transformed this variable and included it in the regression (l2rxc in STATA), but thought we deviated from our research question in doing so, therefore we decided not to use it. Another discarded model included critic reviews, but was excluded for similar reasons, in addition to high correlation with existing predictors and because we lacked information about what the critic reviews entail.

***3.3.2 Final model selection and presentation***

Population equation for the final model:

Fitted equation for the multi-linear regression model:

The y-intercept ( 1.23 represents the predicted log2 gross in dollars for a movie with zero log2 news articles published. The slope coefficient (, .38 is the predicted difference in log2 gross between two action/adventure movies (when act\_adv = 1), or two movies that belong to any genre other than action/adventure (when act\_adv = 0) that differ by one-unit in log2 news articles published, controlling for all other variables in the model. We have convincing evidence that the controlled association is not equal to zero (*t*(237) = 5.42, *p* < .0001), therefore we can reject the null hypothesis and conclude that there is a linear association between news articles published and gross in the larger population of films in the US.

The slope coefficient (, .79, is the predicted difference in log2 gross between two action/adventure movies (i.e. when act\_adv = 1) that differ by one-unit in log2 budget, holding constant all other variables in the model. This association was also significant (*t*(237) = 2.15, *p* < .0001).

The slope coefficient (, .15, is the predicted difference in log2 gross between two action/adventure movies (i.e. when act\_adv = 1) that differ by one-unit in log2 awards received, controlling for all other variables in the model. However, this association was not significant *t*(237) = 1.90, *p* >.05).

The slope coefficient , -.06, is the predicted difference in log2 gross between movies from the action/adventure genres and movies from other genres that are not adventure or action, holding constant all other variables in the model. However, this association was very close to zero and it was not significant *t*(237) = -0.18, *p* >.05).

In terms of variance explanation, this model outperformed other models we tried that are relevant to our research question – its R-squared shows it explains 58.2% of variability in the outcome variable “l2gross”. Fig.4 shows the final regression model and the fit line. We also included two prototypical lines for low- and high-budget movies.

***3.3.3 Regression Diagnostics***

To test the model assumptions required for least squares estimation and inference, we used the standardized residuals of our regression model and plotted it against each of the main predictor, the controls, and the predicted values (yhat).

The assumption of normality is likely violated based on the slight residual left skew in the histogram plot (fig.2) of the standardized residuals. The outlier values of residuals on the plot of (fig.1) also supports this conclusion.

Conditional independence of observations is likely violated; especially because of movies that are part of a series. This dataset does contain several movies that are part of a series (e.g Ironman, Paranormal Activity, Saw, Avengers), so it is reasonable to assume that the sequels are not independent from the prequels; meaning that some moviegoers likely made their decision to see Saw V because they liked Saw IV, for example.

The remainder of this section summarizes the assumptions of heteroscedasticity and linearity per variable (normality and independence are commented on only once as they apply across the model, except for the binary variable).

L2news - our main predictor: Linearity largely holds based on the shape of (fig.1). There do not seem to be regions with systematically positive or negative residuals. Heteroscedasticity seems violated to some extent. Evidence of this can be seen on the right half of the plot (fig.1) which has much lower variability than the left half (with the exceptions of 3 outlier observations).

l2budget: Linearity is clearly violated towards the left end of plot of l2budget on the residuals. However, this is likely due to a few outlier values at that end. It is important to note that, excluding those outliers, the linearity assumption holds for the remainder of the plot. Heteroscedasticity is also violated; where two clear regions display different patterns of variability in residuals.

l2wins: Linearity strongly holds for the standardized residuals on this predictor - this is evident from its fit line and shape of the scatter-plot. Heteroscedasticity is clearly violated since the residuals vary highly at the left end of the plot, then cones down gradually to a region with much less variation at the right end.

act\_adv: Linearity largely holds since the means for observations falling into either category (0,1) are close to zero (.0027 and .0008 respectively). Heteroscedasticity may be violated because the standard deviation for each binary value (0,1) is different - for 0 it is 1.13, while for 1 it is .71.  This difference could indicate a pattern that violates heteroscedasticity, but it is difficult to say for sure due to the binary nature of this predictor and because the SD difference is not very large. Normality is likely violated due to clear evidence of residual skew for the histogram of the (0) and values when this predictor is plotted.

yhatl2gross: Linearity mostly holds for the predicted values of our outcome. This is clear from the shape of the fitted line in (fig.3). There is however two slight bends at the tails of the line. Heteroscedasticity is violated based on the shape of the residual plot in (fig.3). There seems to be a pattern of large variation on the left side; coned down to less variation at the right end.

Overall, we believe that the regression assumptions are somewhat reasonable, and that the model is fairly representative of the controlled relationship despite the violations noted above - namely with heteroscedasticity.

**4. Discussion**

***4.1 Interpretation of results***

The results of our investigation show that there exists a statistically significant (p<.001) association between the number of news articles published about a movie and how much money that movie grosses when controlling for the movie’s budget, number of awards won, and whether it belongs to the genres Action and/or Adventure. The association is upward sloping (positive), linear, and of medium-high strength (red line in fig.4). This is in line with the alternative hypothesis that there is a linear association between our main predictor and outcome variable; controlling for the aforementioned variables. Numerically, the main interpretation of our model is that on average, every doubling in news coverage (l2news) for a given movie is associated with a .38 doubling of that movie’s gross in USD (l2gross). Coming as no surprise, when we introduced the prototypical lines for low- and high- budget movies (10th and 90th percentiles respectively), the low-budget movies had a lower y-intercept than the high-budget ones (fig.4).

Our model also has a relatively high explanatory power of the variation in the predicted outcome, based on the R-squared measure of .582; meaning that the predictors used explain 58.2% of the variation in l2gross (table.3).

***4.2 Implications***

The main implication of this model is that news coverage is linked to how much a movie grosses. This information is likely interesting to members of the film industry, or producers as they plan marketing for their movies. Our study provides empirical evidence of this association when controlling for other select variables. While our study did not assess the causation or draw a distinction between the type of news articles required to have this impact on gross revenue, it seems that – as far as our sample goes – more news coverage in general is associated with a movie making more money.

***4.3 Limitations***

Most obviously, this study establishes an association, but does not look into causation or the direction of this causation. As such, it is not clear whether news coverage influences movie gross or whether movie gross influences news coverage. Further to this point, the direction of causation may be a mix of both directions depending on the time since release, for example, among other potential variables. This could be studied through time-series analysis.

A second drawback is the sample size. 339 observations from the USA from 2005 onward are arguably insufficient to draw an inference that explains a relationship over a wider population of movies. This sample fell to 242 with log2 transformation of “wins” variable, as many movies had 0 awards.

Third, the gross figures used are only from the USA and in USD. In reality, American movies are watched by many people outside the USA too, and success internationally could be an important factor for producers to consider. As such, this shortcoming limits this research to considering the investigated association inside of the USA only – this association could turn out to be very different depending on the country.

Fourth, is the number of observations eliminated during data cleaning. The used observations are about 2% of the original datasets. Most eliminations were due to missing values.

Lastly, we did not consider the different types of news articles. Such information was absent from the dataset, but it would be interesting to dichotomize the variable “news” into favorable and unfavorable coverage; then include that as a binary variable in our model.

***4.4 Future directions***

Broadly, future research should include a larger dataset that addresses the shortcomings mentioned earlier. We think it is important to investigate this association while taking the type of news coverage into account, not merely the overall coverage as we did in this study. Further, time-series analysis should be incorporated to better understand the nature of this association. Lastly, future researchers are encouraged to apply subject matter expertise to posit hypotheses and explore different variable interactions and controls beyond those applied in this study.

***4.5 Conclusions***

To conclude, news coverage is associated with movie gross income. This study explored this relationship controlling for variables of interest and found a statistically significant association. We iterated different models before settling on our selected one; which had a high explanatory power of the variation in the outcome, among other relevant aspects. The assumption of heteroscedasticity in our model seems violated to some extent. We believe that this model is representative of a very specific movie segment, and we believe further investigation is necessary to apply it outside the discussed context.

**5. References**

Nithin, V. R., Pranav, M., Sarath, B., & Lijiya, A. (2014). Predicting Movie Success Based on IMDB Data. *International Journal of Data Mining Techniques and Applications*, *3*, 365-368.

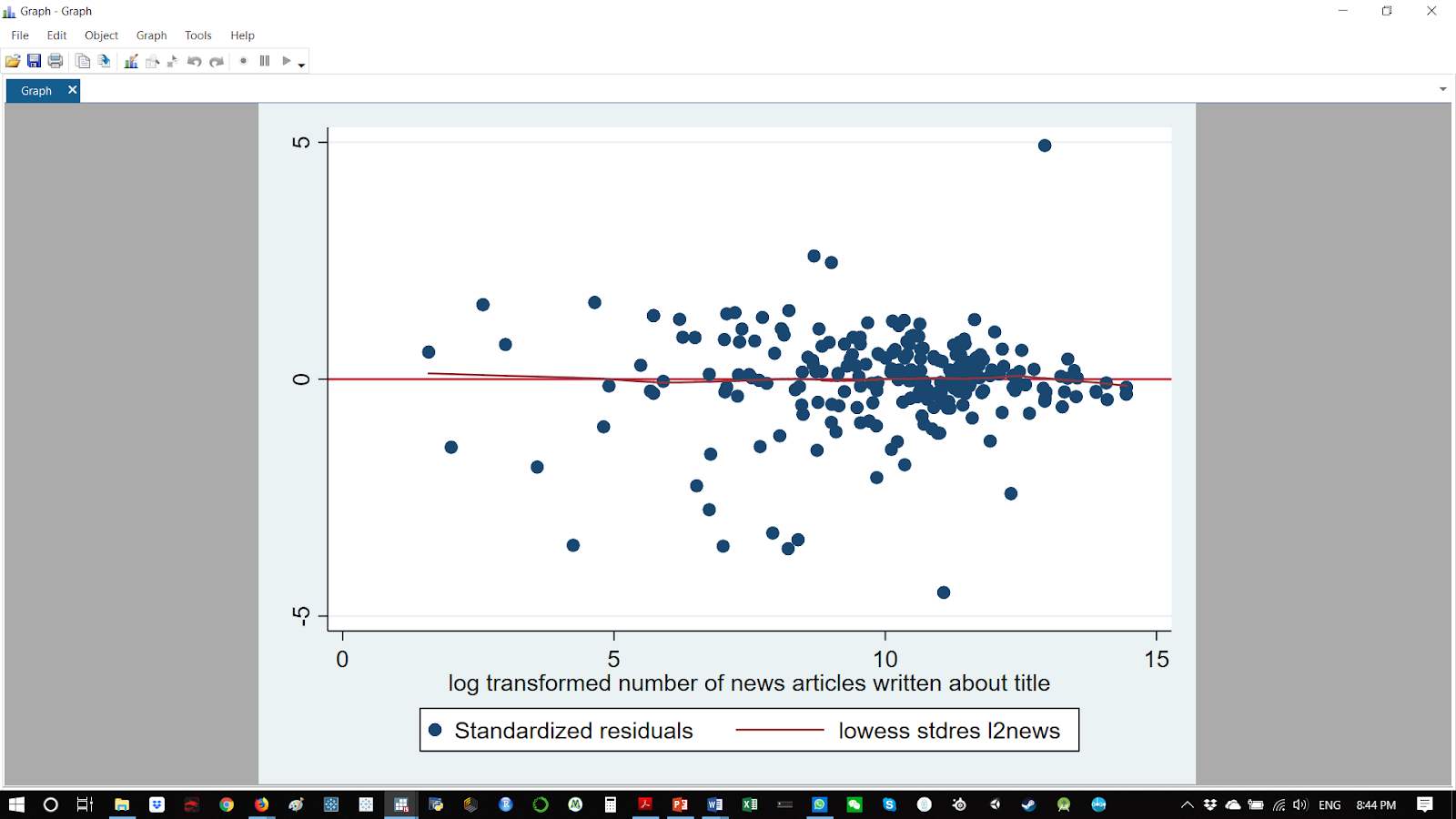
**6. Appendix**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 1. Descriptive Statistics for Continuous Variables** | | | | |
|  | Mean | Median | Standard Deviation | Inter-quartile Range |
| Movie gross |  |  |  |  |
| News articles | 1957.51 | 924 | 3054.99 | 2035 |
| Movie budget |  |  |  |  |
| Number of awards | 9.40 | 2 | 18.42 | 9 |
| **Descriptive Statistics for continuous variables after log-transformation** | | | | |
| Movie gross (*log2*) | 24.32 | 25.07 | 3.11 | 3.08 |
| News articles (*log2*) | 9.56 | 9.86 | 2.38 | 2.85 |
| Movie budget (*log2*) | 24.29 | 24.58 | 2.32 | 2.5 |
| Number of awards (*log2*) | 2.33 | 2 | 1.98 | 2.91 |
| *N* | 339 | 339 | 339 | 339 |
| Note: Mean, Median, SD, and IQR for Movie gross and Movie Budget are written in scientific notation due to its large size. | | | | |

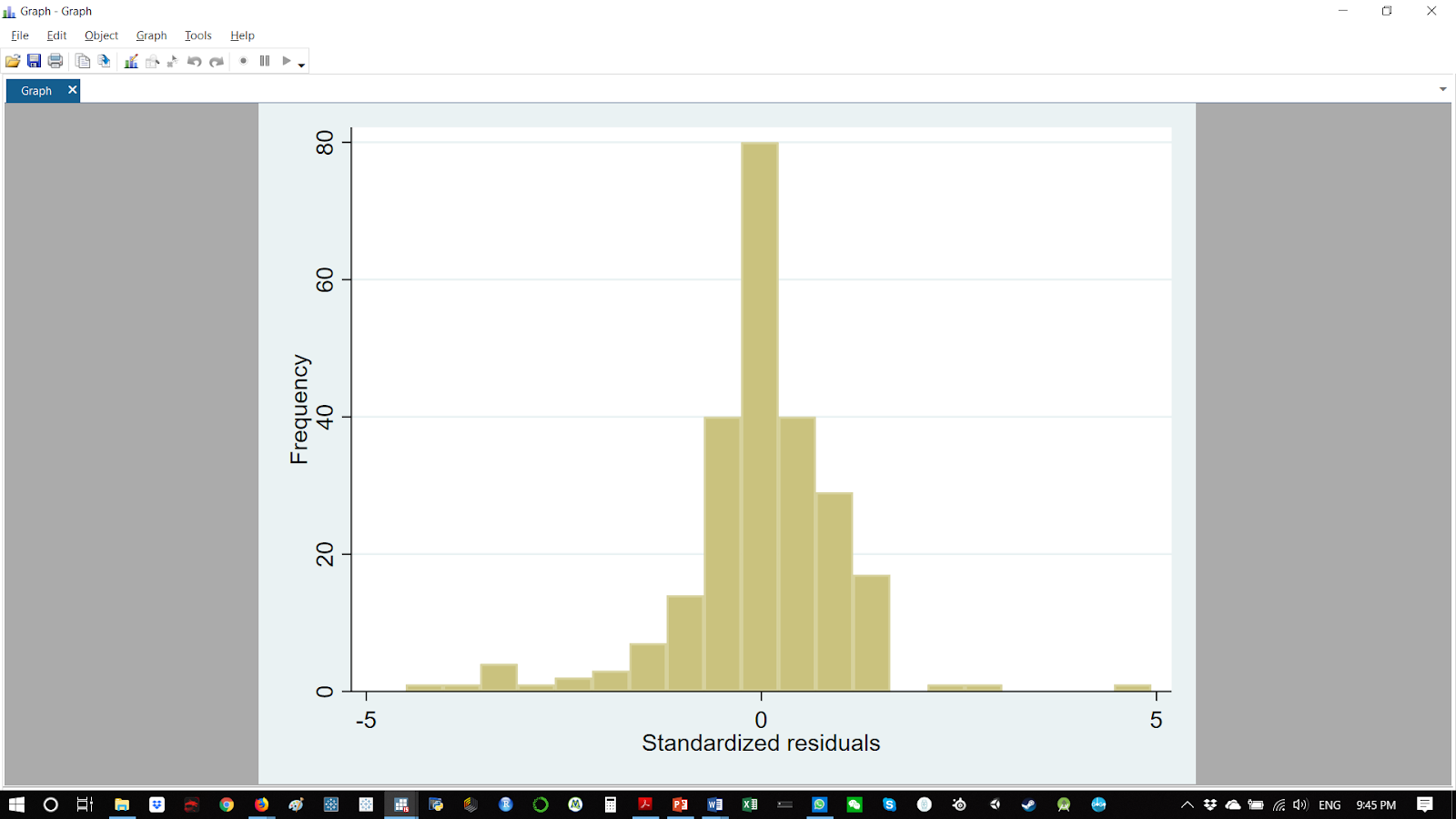
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 2. Correlation matrix (n = 435)** | | | | |  |
|  | Movie gross  (*log2*) | News articles (*log2*) | Movie budget  (*log2*) | Number of awards (*log2*) | Action-  adventure genre |
| Movie gross (*log2*) | 1.00 |  |  |  |  |
| News articles (*log2*) |  | 1.00 |  |  |  |
| Movie budget (*log2*) |  |  | 1.00 |  |  |
| Number of awards (*log2*) |  |  | 0.03 | 1.00 |  |
| Action-adventure genre |  |  |  | -0.09 | 1.00 |
| \**p <.05* | | | | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 3. Taxonomy Table of Regression Models** | | | | |
|  | (final model) | (3rd attempt) | (2nd attempt) | (initial model) |
|  | l2gross | l2gross | l2gross | gross |
|  | b/se/t | b/se/t | b/se/t | b/se/t |
| l2news | 0.38\*\*\* | 0.52\*\*\* | 0.73\*\*\* |  |
|  | (0.07) | (0.05) | (0.06) |  |
|  | 5.42 | 9.55 | 12.16 |  |
| l2budget | 0.79\*\*\* | 0.64\*\*\* |  |  |
|  | (0.06) | (0.06) |  |  |
|  | 12.15 | 10.73 |  |  |
| l2wins | 0.15 |  |  |  |
|  | (0.08) |  |  |  |
|  | 1.90 |  |  |  |
| act\_adv | -0.06 | -0.07 |  |  |
|  | (0.33) | (0.29) |  |  |
|  | -0.18 | -0.25 |  |  |
| news\_articles |  |  |  | 17479.38\*\*\* |
|  |  |  |  | (1004.28) |
|  |  |  |  | 17.40 |
| \_cons | 1.23 | 3.76\*\* | 17.38\*\*\* | 29402808.89\*\*\* |
|  | (1.51) | (1.38) | (0.59) | (3640031.90) |
|  | 0.81 | 2.72 | 29.56 | 8.08 |
| *N* | 242 | 337 | 337 | 339 |
| r2 | 0.582 | 0.506 | 0.306 | 0.473 |
| F | 82.51 | 113.7 | 147.9 | 302.9 |
| df\_m | 4 | 3 | 1 | 1 |
| df\_r | 237 | 333 | 335 | 337 |
| rmse | 2.044 | 2.207 | 2.607 | 56405362.6 |
| p | 8.92e-44 | 1.05e-50 | 1.93e-28 | 7.43e-49 |
| \* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001 | | | | |
|  |  |  |  |  |

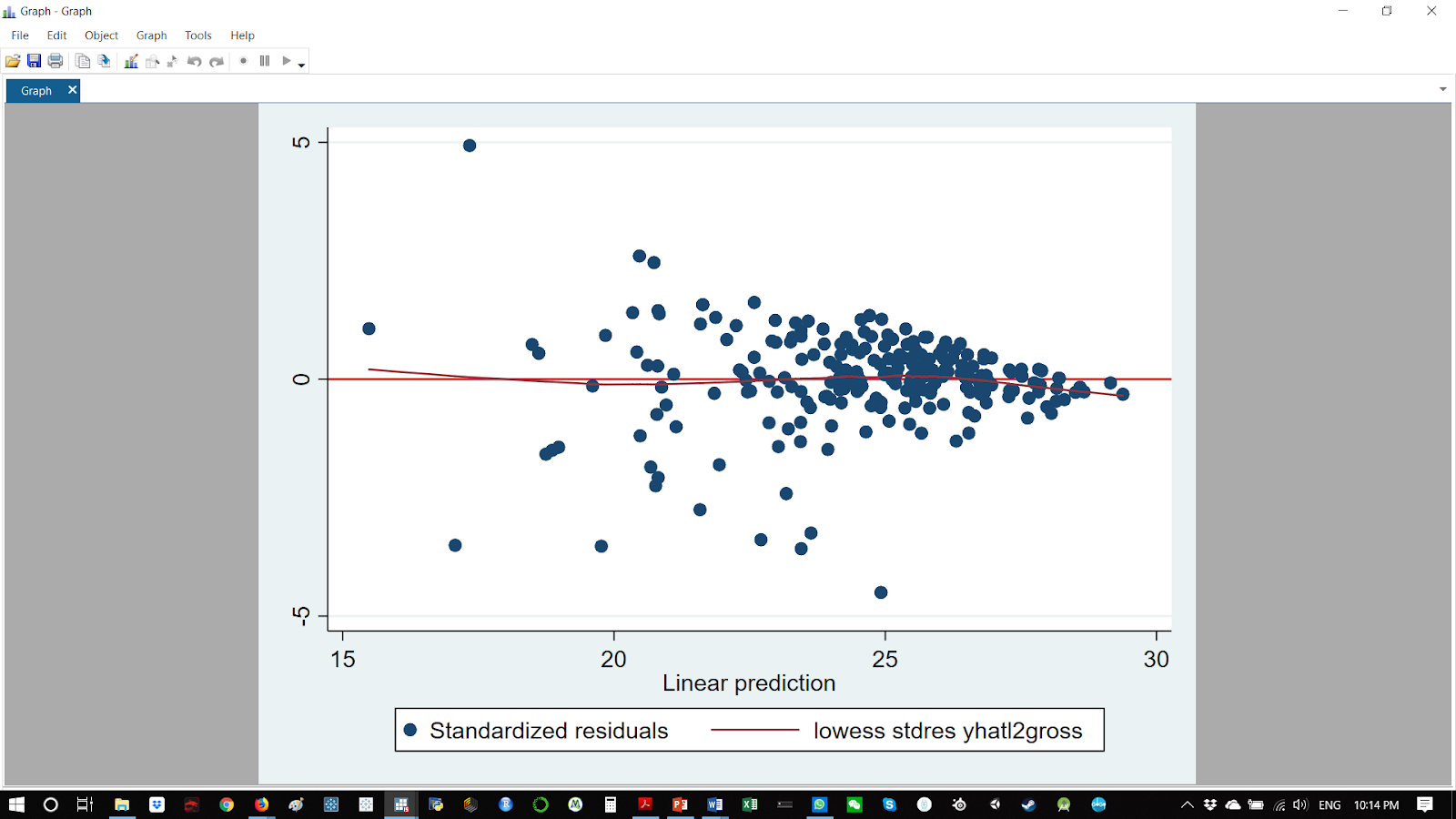
**Fig.1 plot of residuals on main predictor (l2news)**



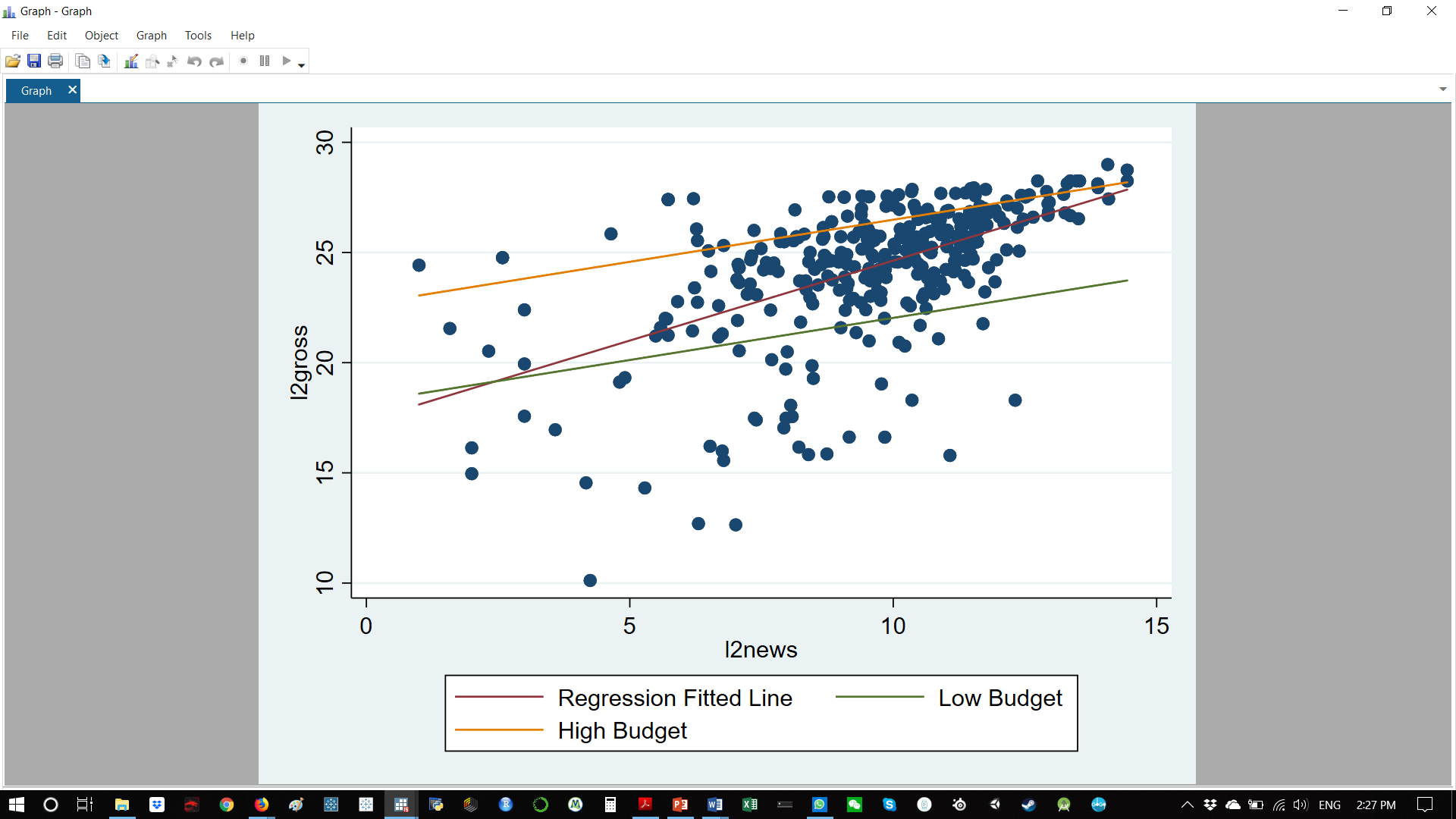
**Fig.2 Histogram of std. residuals of final regression model to spot residual skew**



**Fig.3 plot of residuals on predicted values (yhatl2gross)**

****

**Fig.4 controlled association of final model, line of best fit, and prototypical lines for high- and low-budget movies**



**Annotated STATA code**

// please make sure that the 2 .csv source files and the STATA .do

// file are in the same folder on your computer for the code to run properly

// Change the 4 paths highlighted in yellow to the file path on your computer

////////////////////////////////////////////////////////////////////////

// 1. Cleaning imdb.csv (master file)

clear

cd "C:\Users\Raouf\Desktop\S040\Final"

import delimited imdb.csv

// keeping only type == movies (not episodes or tv series)

drop if type != "video.movie"

// dropping unrequired columns

drop url type fn tid title

// drop missing values

drop v45 v46 v47 v48

foreach v of var \* {

drop if missing(`v')

}

// renaming title column

rename wordsintitle new\_title

// capitalizing each word of movie titles to ease integration with second file

replace new\_title = proper(new\_title)

// saving as .dta

save imdb.dta, replace

////////////////////////////////////////////////////////////////////////

// 2. Cleaning popculture-imdb-5000-movie-dataset.csv (second file)

clear

cd "C:\Users\Raouf\Desktop\S040\Final"

import delimited popculture-imdb-5000-movie-dataset.csv

// keeping only required variables

keep num\_critic\_for\_reviews gross movie\_title num\_voted\_users facenumber\_in\_poster country budget

// drop missing values

foreach v of var \* {

drop if missing(`v')

}

// removing text error in "movie\_title" names

replace movie\_title = substr(movie\_title,1,length(movie\_title)-4)

// capitalizing each word of movie titles to ease integration with master file

replace movie\_title = proper(movie\_title)

rename movie\_title new\_title

// saving as .dta

save popculture-imdb-5000-movie-dataset.dta, replace

////////////////////////////////////////////////////////////////////////

// 3. combining both files by "new\_title"

clear

cd "C:\Users\Raouf\Desktop\S040\Final"

local myfilelist : dir . files "\*.dta"

use imdb

merge m:m new\_title using popculture-imdb-5000-movie-dataset.dta

drop \_merge

save imdb\_combine, replace

////////////////////////////////////////////////////////////////////////

// 4. Using the combined file to conduct analysis

clear

cd "C:\Users\Raouf\Desktop\S040\Final"

use imdb\_combine.dta

// removing all NA values

foreach v of var \* {

drop if missing(`v')

}

// dropping moives outside of the USA because some figures were reported in non USD currencies

drop if country != "USA"

drop country

// converting string variables to numeric

destring imdbrating nrofwins nrofnominations nrofphotos nrofnewsarticles nrofuserreviews nrofgenre duration ratingcount, replace

// dropping redundant variable between two datasets

drop num\_voted\_users

// renaming long variable names for convenience

rename nrofwins wins

rename nrofnominations nominations

rename nrofphotos photos

rename nrofnewsarticles news\_articles

rename nrofuserreviews reviews

rename nrofgenre genres

rename num\_critic\_for\_reviews critic

rename facenumber\_in\_poster faces

// adding labels

label variable duration "run time in minutes"

label variable critic "number of critic reviews"

label variable faces "number of faces on official movie poster"

label variable year "year movie released"

// converting year from string to numeric then dropping year<2005

destring year, replace

drop if year <2005

// exploring all variables by regressing everything

regress imdbrating ratingcount duration wins nominations photos news\_articles reviews genre action adult adventure animation biography comedy crime documentary drama family fantasy filmnoir gameshow history horror music musical mystery news realitytv romance scifi short sport talkshow thriller war western critic gross faces budget

// dropping vairables identified as collinear by Stata

drop adult filmnoir gameshow news realitytv talkshow

// exploring gross and budget

summ budget gross, detail

twoway (scatter gross budget,yline(0))

// generating a profit variable

generate profit = gross - budget

label variable profit "gross - budget"

// log transforming for better representation

generate l2gross = log(gross)/log(2)

generate l2budget = log(budget)/log(2)

generate l2reviews = log(reviews)/log(2)

generate l2wins = log(wins)/log(2)

label variable l2wins "log transformed number of awards won"

generate l2news = log(news\_articles)/log(2)

label variable l2news "log transformed number of news articles written about title"

// generating a composite of action or adventure

generate act\_adv = 0

replace act\_adv = 1 if action==1 | adventure==1

label variable act\_adv "binary variable for either action or adventure (0= not action nor adventure, 1= is action or adventure or both)"

browse action adventure act\_adv

// examining the plot of L2gross L2budget

twoway (scatter l2gross l2budget,yline(0))(lfit l2gross l2budget)

// R.1 for low(25%) 1 and high (75%)4 l2wins

regress l2gross l2news l2budget l2wins act\_adv

twoway (scatter l2gross l2news)(function y = \_b[\_cons] + \_b[l2news]\*x + \_b[l2budget]\*24.2 + \_b[l2wins]\*1 + \_b[act\_adv]\*1, range(l2news))(function y = \_b[\_cons] + \_b[l2news]\*x + \_b[l2budget]\*24.2 + \_b[l2wins]\*3.81 + \_b[act\_adv]\*1, range(l2news)), legend(order(2 "Low wins" 3 "high wins"))ytitle("l2gross")xtitle("l2news")

// R.2 including critic (25% and 75%iles)

regress l2gross l2news l2budget l2wins critic act\_adv

twoway (scatter l2gross l2news)(function y = \_b[\_cons] + \_b[l2news]\*x + \_b[l2budget]\*24.2 + \_b[l2wins]\*2.4 + \_b[critic]\*100 + \_b[act\_adv]\*1, range(l2news))(function y = \_b[\_cons] + \_b[l2news]\*x + \_b[l2budget]\*24.2 + \_b[l2wins]\*2.4 + \_b[critic]\*267 + \_b[act\_adv]\*1, range(l2news)), legend(order(2 "Low wins" 3 "high wins"))ytitle("l2gross")xtitle("l2news")

// R.3 introducing and experimenting with l2rxc

generate rxc = imdbrating \* ratingcount

label variable rxc "ratingcount x imdbrating"

generate l2rxc = log(rxc)/log(2)

label variable l2rxc "log2 of ratingcount x imdbrating"

regress l2gross l2news l2budget l2rxc act\_adv

twoway (scatter l2gross l2news)(function y = \_b[\_cons] + \_b[l2news]\*x + \_b[l2budget]\*24.2 + \_b[l2rxc]\*17.08 + \_b[act\_adv]\*1, range(l2news))(function y = \_b[\_cons] + \_b[l2news]\*x + \_b[l2budget]\*24.2 + \_b[l2rxc]\*17.08 + \_b[act\_adv]\*1, range(l2news)), legend(order(2 "Low l2rxc" 3 "high l2rxc"))ytitle("l2gross")xtitle("l2news")

// R.4 including and varying critic (25% and 75%iles) instead of l2rxc

regress l2gross l2news l2budget l2rxc critic act\_adv

twoway (scatter l2gross l2news)(function y = \_b[\_cons] + \_b[l2news]\*x + \_b[l2budget]\*24.2 + \_b[l2rxc]\*18.21 + \_b[critic]\*100 + \_b[act\_adv]\*1, range(l2news))(function y = \_b[\_cons] + \_b[l2news]\*x + \_b[l2budget]\*24.2 + \_b[l2rxc]\*18.21 + \_b[critic]\*267 + \_b[act\_adv]\*1, range(l2news)), legend(order(2 "Low critic" 3 "high critic"))ytitle("l2gross")xtitle("l2news")

// R.5 introducing controversial

//generate controversy = critic \* news

//label variable controversy "interaction between news and critic used to proxy for how controversial a movie is"

//generate l2cnvs = log(controversy)/log(2)

//label variable l2cnvs = log2 "transformation of controversy"

sum gross if action ==1

sum gross if adventure ==1

// Exploring correlations

pwcorr l2gross duration photos genres action adventure animation biography comedy crime documentary drama family fantasy history horror music musical mystery romance scifi short sport thriller war western critic faces l2budget l2reviews l2wins l2news rxc

// correlation matrix

pwcorr l2news l2budget l2wins critic act\_adv, star(0.05) sig

// Correlation matrix graph

graph matrix l2gross l2news l2budget l2wins act\_adv

// producing taxonomy table

eststo: regress l2gross l2news l2budget l2wins act\_adv

est sto M1

eststo: regress l2gross l2news l2budget act\_adv

est sto M2

eststo: regress l2gross l2news

est sto M3

eststo: regress gross news

est sto M4

esttab M1 M2 M3 M4 using "taxonomy\_table2.rtf", replace cells(b(star fmt(2)) se(par fmt(2)) t(fmt(2))) scalars(r2 F df\_m df\_r rmse p) legend

// residuals and assumptions testing on final model predictors and yhat

regress l2gross l2news l2budget l2wins act\_adv

predict yhatl2gross, xb

predict stdres, rstandard

twoway (scatter stdres l2news, yline(0))(lowess stdres l2news)

twoway (scatter stdres l2budget, yline(0))(lowess stdres l2budget)

twoway (scatter stdres l2wins, yline(0))(lowess stdres l2wins)

twoway (scatter stdres act\_adv, yline(0))(lowess stdres act\_adv)

twoway (scatter stdres yhatl2gross, yline(0))(lowess stdres yhatl2gross)

hist stdres

// examining act\_adv

bysort act\_adv: summarize stdres

hist stdres, by(act\_adv) bin(100)

// Including prototypical lines for act\_adv low budget (10th pctl) and high budget (90th pctl) movies

regress l2gross l2news l2budget l2wins act\_adv

twoway (scatter l2gross l2news)(lfit l2gross l2news)(function y = \_b[\_cons] + \_b[l2news]\*x + \_b[l2budget]\*21.25 + \_b[l2wins]\*2 + \_b[act\_adv]\*1, range(l2news))(function y = \_b[\_cons] + \_b[l2news]\*x + \_b[l2budget]\*26.90 + \_b[l2wins]\*2 + \_b[act\_adv]\*1, range(l2news)), legend(order(2 "Regression Fitted Line" 3 "Low Budget" 4 "High Budget"))ytitle("l2gross")xtitle("l2news")

1. https://www.kaggle.com/orgesleka/imdbmovies/home [↑](#footnote-ref-1)
2. https://data.world/popculture/imdb-5000-movie-dataset [↑](#footnote-ref-2)