Statistics Capstone Kickstarter Project Report

for

Dr. Lan Zhou
Associate Professor
Texas A&M University
College Station, TX

by

Jacy Dong
Eric-Steven Hanson
Caroline Lee
Ian Scarff
STAT 482-900
May 3, 2019

Introduction

Purpose

Kickstarter is a crowdfunding platform that brings projects to life by allowing consumers to contribute money to projects they're interested in. The creator of a Kickstarter sets their project's funding goal and deadline, and if people like the project, they pledge money to said project. If the project succeeds by reaching at least 100% of its funding goal, all backers are charged for their monetary contribution. Kickstarter operates with a "all or nothing system", and if the project falls short of its funding goal, no one is charged.

Kickstarter has launched itself into the mainstream due to its most prolific successes and failures. The platform was the birthplace of projects such as the Fidget Cube and Exploding Kittens, a card game. As of the April 2019, 161,937 projects have been funded with over 4 trillion total USD. Despite the large number of Kickstarter successes, there are even more projects that have failed to reach their funding goals. According to Kickstarter, only 36.92% of the projects on their platform have been successful. However, it's not immediately clear *why* some projects succeed while others fail.

The aim of this project is to determine what factors are important for determining the success of a Kickstarter and which factors have an effect on the percentage of the goal a Kickstarter reaches. This study is important for people who are currently in or who want to start a Kickstarter. Our findings will inform Kickstarter users if their projects are predicted to fail or not, allowing them to make adjustments if necessary and what adjustments they should make in order to reach their goals.

Data Overview

We used Kickstarter data from the Kickstarter Projects page on Kaggle.com. The dataset contains over 379,000 different Kickstarter projects between 2009 and 2018 with 13 variables per project. Although the projects in this dataset come from many different countries, we will only use Kickstarters originating from the United States. This removes any undefined variables and allows us to ignore the variables country, currency, and the USD conversions.

There are 292,627 Kickstarters that originated from the US. From this group, 269,071 (92%) Kickstarters have monetary goals ranging from $100 to $50,000. We believe that projects with this monetary goal range represent the most common and legitimate projects that will produce finished products, and not people who are trying to make a quick buck or have extreme, unrealistic monetary goals. As indicated by Figure 1 and Table 1, most projects on Kickstarter are in this monetary goal range. Furthermore, we will only use Kickstarters that have the state of failure, success or live. Kickstarters that have the state of canceled or suspended have too many possible outside reasons for their "failure." We want to solely focus on whether or not a Kickstarter makes enough money.

Table 1.

	Min	Max	Mean	Median
All US Data	$0.01	$100 M	$44,036.75	$5,250
Subset US Data	$100	$50,000	$9,682.21	$5,000

From this data, the variables of interest are: 1) the main category of the project, 2) various subcategories for each main category, 3) the month a project was created, 4) a monetary goal, 5) the year a project was created, 6) how much money was pledged, 7) the current state of the project, and 8) the number of backers.

Hypotheses

Using this data, we will test the following hypotheses:

1. The runtime of a project has a positive effect on the success of a project.
2. The size of the monetary goal has a negative effect on the success of a project.
3. The size of the monetary goal has a negative effect on the amount of money raised.
4. The type of project has an effect on the success of a project.
5. The type of project has an effect on the amount of money raised.

Methods

To test these hypotheses, we developed two models: a logistic regression model for predicting the success of a Kickstarter, and a linear regression model for predicting the amount of money raised. Power transformations were used in the construction of each model. Furthermore, since we were interested in determining the probability of success for a Kickstarter based on the logistic regression model, we used the Chi-Square test to choose the best logistic regression model. Similarly, we used the Likelihood Ratio test to choose the best linear regression model. To validate model predictions, the models were compared against a test dataset.

To build and our models, we took a random sample of 10,000 U.S. projects with their state being either success or failure. This sample was randomly split into a training dataset containing approximately 75% of the sample, while the remaining 25% comprised the test dataset. We will used training data to build our models.

In addition to the variables stated above, we created five more variables. The first variable, denoted as *RunTime*, is the duration of a Kickstarter in days. The second variable, denoted as *TimeDiff*, is the difference in the number of days between the first project created on Kickstarter to the project in question. The third variable, denoted as *PercentGoal*, is the percentage of goal reached. The fourth and fifth variables, denoted as *Month* and *Year* respectively, are the month and year a Kickstarter was launched. Both will be treated as categorical variables. The month a Kickstarter is launched could potentially affect its success because certain months could have holidays that influence potential backers to spend their money elsewhere. The year a Kickstarter is launched could also have an effect because the amount of people of using Kickstarter has increased and the public perception of the website has changed over the years.

Exploratory Data Analysis

For this data analysis, we used our random sample of 10,000 projects. We generated a scatterplot matrix to observe the relationships between the continuous variables in our model and color coded based on the projects state. Based on these plots, we found that the data is unorganized, follows no specific patterns, and contains many outliers. However, the RunTime and TimeDiff variables were decently centered on their respective medians. To combat skewness, we applied a log transformation to the remaining variables. We accounted for the zeros in our data by adding the appropriate minimum value based on the variable. For backers, the minimum is 1; for pledged, the minimum is $0.01; and for percent pledged, we decided on an arbitrary minimum of 0.01%.

Figure 2.

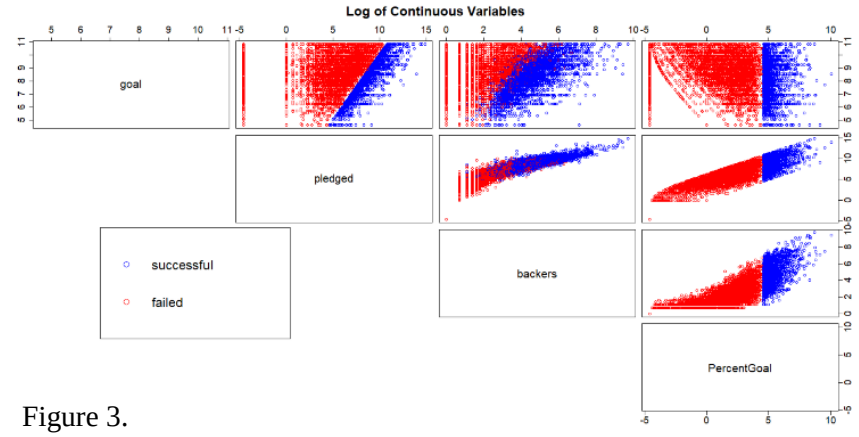
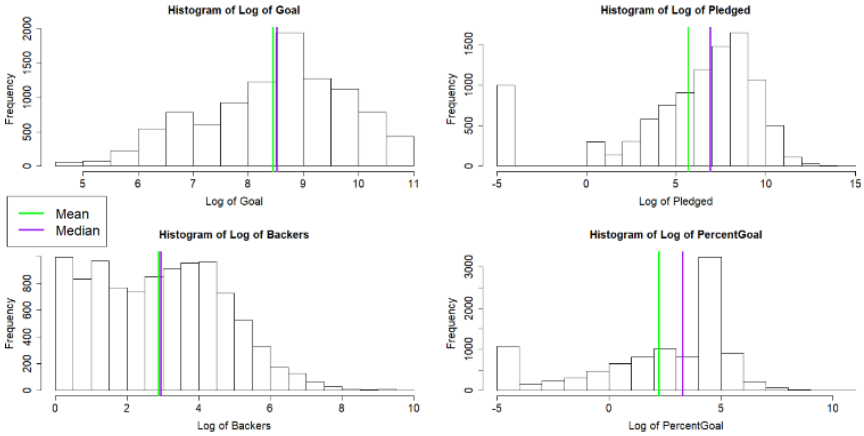


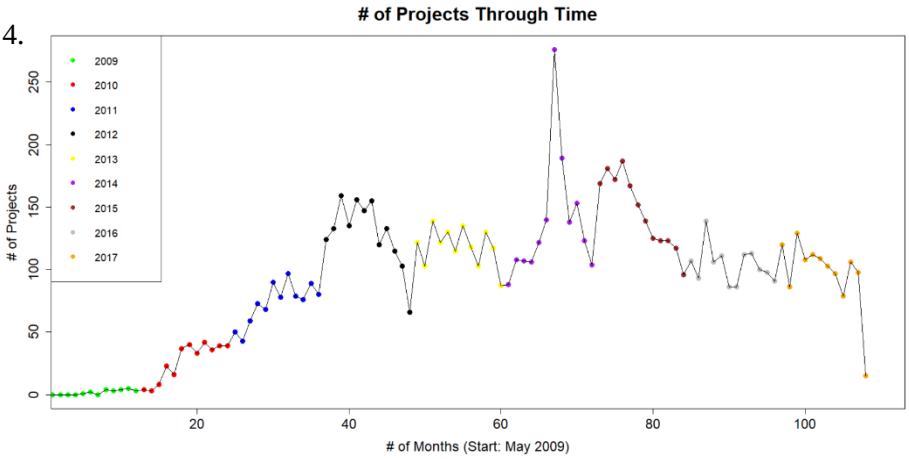
Figure 3.



The result of this is shown in the Figures 2 and 3 to the left. Based on Figure 2, the log transformation has brought in outliers and has established some linear relationships. We can also see a possible dividing line between a successful project and a failed project, which is color coded in blue and red, respectively. This may be useful for future modeling and predicting. However, in Figure 3, we can still see that there is some skewness, especially in the number of backers. Further transformations will be tested during the development of models.

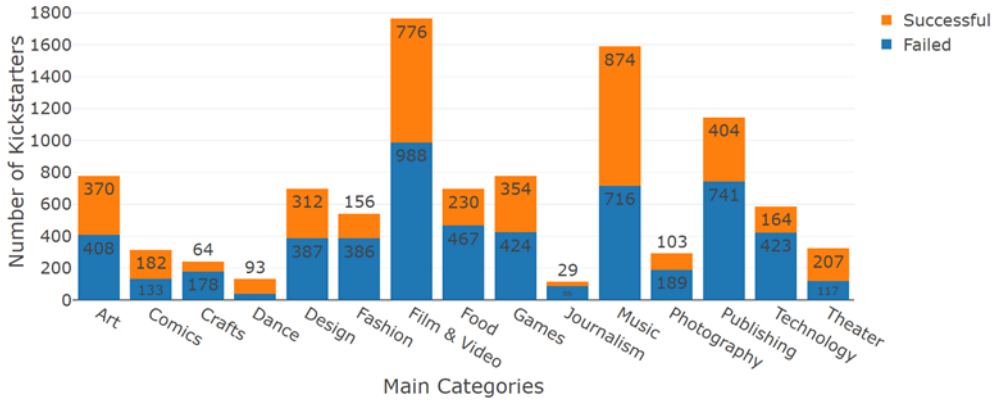
Next, we examined how the number of projects launched changed over time. From Figure 4 below, we can see that after the website was launched in April 2009, it was slow to gain popularity. However, starting in 2010, the number of projects created grew fast, peaking in July 2014. Thereafter, the number returned to the same levels as in 2012 and 2013. Since then, the popularity of Kickstarter seems to have dropped, sharply in December 2017.

Figure 4.



Next, we examined how the data was distributed across each of the main categories. We first examined the number of Kickstarter in each category, as displayed in Figure 5 below. In our random sample, Film & Video, Music, and Publishing are the three most popular categories. However, we can see that smaller categories, such as Comics, Dance, and Theater have high success rates.

Figure 5.



We then examined the distribution of the number of backers and the amount pledged. As expected, these variables were heavily skewed. To correct for skewness, we shifted them by their respective minimums and applied a log transformation. As shown in Figures 6 and 7 below, they are now more centered on their median. From these plots, we can see that the median for log(backers) is about the same through all main categories. However, the variances in each category do vary. For example, the variance of log(backers) under the Food category is different

Figure 6.

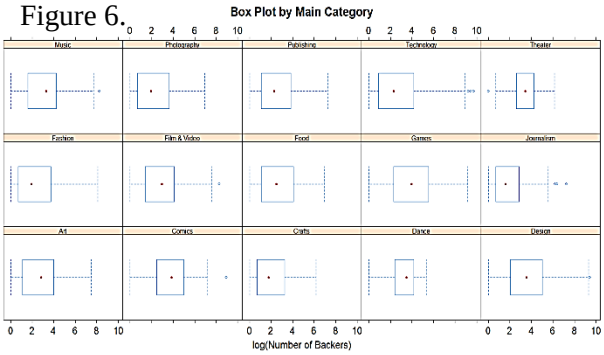
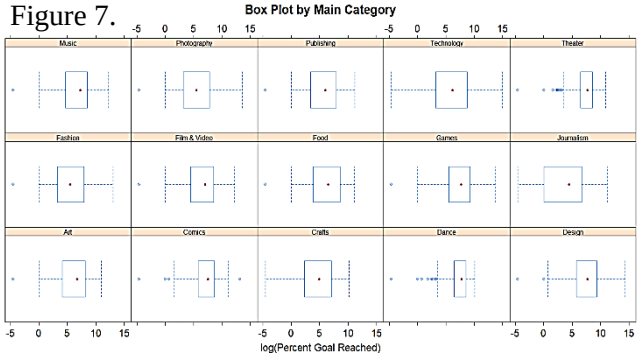


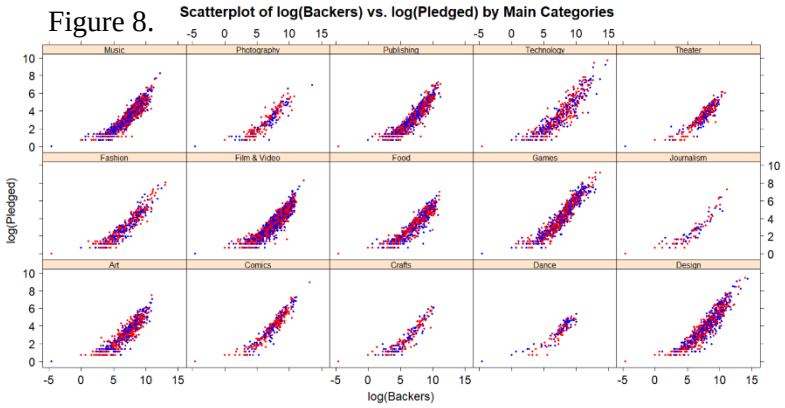
Figure 7.



than the variance under the Technology category. The same story is told for log(Percent Goal Reached).

Finally, we plotted the log of backers by the log of the amount pledged, separated by categories, displayed in Figure 8 below. We can see that there may be small quadratic relationships between these variables in each category. However, there is no clear distinction what values determine a successful and an unsuccessful project, which are color coded blue and red, respectively. The addition of subcategories to each main category will aid in the determination of success and failure. This will be determined while developing models.

Figure 8.



Model Building

Logistic Regression Model

For our first statistical model, we constructed a logistic regression model to predict a Kickstarter's probability of success. This model was used to test hypotheses 1, 2, and 4. The data used to build the model comes from the training dataset that was stated earlier in this report.

Based on our exploratory analysis, we explored various power transformations for the continuous variables. To do this, we used the powerTransform function from the R package "car." To apply this function, we first shifted both backers and TimeDiff by 1. Table 2 to the left shows Box Cox power transformation to establish multinormality. The powerTransform function also computed Likelihood Ratio tests for $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 0$ and $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 1$. The p-value for the first test was less than 1e-10, signifying that applying a log transformation to every variable is not recommended. However, the p-value for the second test was also less than 1e-10, signifying that some power transformation should be applied. Based on this table, we applied a logarithmic transformation to goal and backers, and a square root transformation to RunTime. No transformation was recommended for TimeDiff.

Table 2.	Est. Power	Rounded Pwr
Goal	0.1101	0.11
RunTime	0.5052	0.5
Backers	-0.1094	-0.11
TimeDiff	0.921	0.92

A logistic regression model with explanatory variables log(goal), RunTime$^{0.5}$, log(backers), TimeDiff, Month (treated as a factor with 12 levels), Year (treated as a factor with 9 levels), and Subcategory (treated as factor variable with 159 levels). To test the possibility that TimeDiff might explain month and year, we used the Chi-Square test to select variables. We compared models based on deviance and AIC. Our final model contains the variables log(goal), √RunTime, log(backers), TimeDiff, and Subcategory (treated as factor variable with 159 levels). Table 3 displays the estimated log odds of the continuous variables.

		95% Confidence Interval	
	Estimate	Lower Bound	Upper Bound
log(goal)	-2.191	-2.238	-2.145
√Runtime	-0.059	-0.093	-0.024
log(backers)	3.44	3.380	3.502
TimeDiff*	8.310	3.057	13.569

*Note: Multiply by 1e-05.

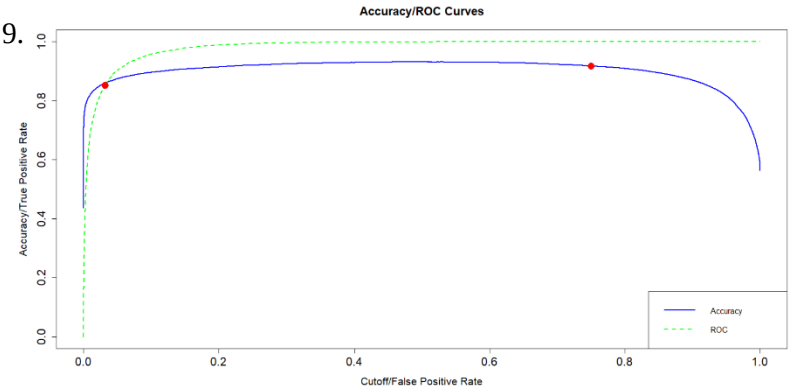
From Table 3, we can see that increases in either goal or run time have negative effects of -2.191 (p-value < 1e-10) and -0.059 (p-value < 0.001), respectively, on the probability of success and that increases in either backers or TimeDiff have positive effects of 3.44 (p-value < 1e-10) and 8.310* (p-value < 0.002). The type of project also has an effect on the probability of success. When compared to the base category 3D Printing, some projects will have an advantage and some will have a disadvantage. For example, a project that will produce Tabletop Games has an estimated negative effect of -2.914 with a 95% confidence interval of (-3.947, -1.881) and a p-value of 1.69e-08 and a project that will produce Performances has an estimated positive effect of 9.378e-01 with a 95% confidence interval of (-0.229, 2.104), but with a p-value of 0.108.

To test our models, we made predictions across our test dataset. By examining accuracy and ROC curves, we set a success value at 75%. This cutoff is realistic in determining the success of a Kickstarter. Table 4 to the left is a confusion matrix with an accuracy measurement. From this table, the model has a high accuracy. The accuracy and ROC curves are shown below in Figure 9, with our model predictions at a 75% cutoff marked by a red dot. Based on this graph, a cutoff of 75% provides high accuracy and a good true/false positive rate tradeoff.

	Test Data	
	Fail	Success
Pre.Fail	13639	1625
Pre.Sucess	452	9284
Accuracy	91.69%	

Table 4.

Figure 9.



Linear Regression Model

For our second statistical model, we constructed a linear regression model to predict the log of the amount of money pledged to a Kickstarter. This model was used to test hypotheses 3 and 5. The data used to build the model comes from the training dataset that was stated earlier in this report.

As in our first model, we used the powerTransform function from the R package "car" to determine necessary power transformations. As before, we applied logarithmic transformations to goal and backers, a square root transformation to RunTime, and no transformation to TimeDiff. In addition, we used the powerTransform function to determine a transformation for pledged. From this, a logarithmic transformation was applied to pledged. A linear regression model with explanatory variables log(goal), RunTime0.5, log(backers), TimeDiff, Month (treated as a factor with 12 levels), Year (treated as a factor with 9 levels), and Subcategory (treated as factor variable with 159 levels) and response variable log(pledged) was fitted to the training dataset. We used the Likelihood Ratio test for the removal of variables. After testing various reduced models, TimeDiff was the only variable removed. Our final model has an R^2 of 0.983 and contains the explanatory variables log(goal), RunTime0.5, log(backers), TimeDiff, and Subcategory (treated as factor variable with 159 levels).

Table 5 displays the estimated effects of the continuous variables. From this table, we can see that increases in either goal (in log($)) or backers (in log(persons)) have positive effects of 0.132 (p-value < 1e-10) and 1.267 (p-value < 5e-7), respectively, on amount of money pledged and that increases in RunTime (in square root of days) have a negative effect of -0.019 (p-value < 1e-10). The month, year, and subcategory of the Kickstarter had effects on the amount of money pledged. Figures 10, 11, and 12 summarize the estimates and 95% confidence intervals for each variable.

Table 5.

		95% Confidence Interval	
	Estimate	Lower Bound	Upper Bound
log(goal)	0.132	0.126	0.138
$Runtime^{0.5}$	-0.019	-0.026	-0.012
log(backers)	1.267	1.263	1.271

Figure 10.

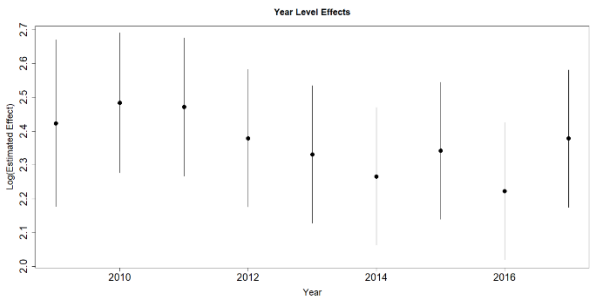


Figure 11.

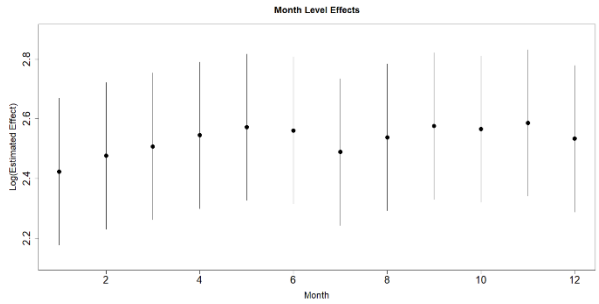


Figure 12.

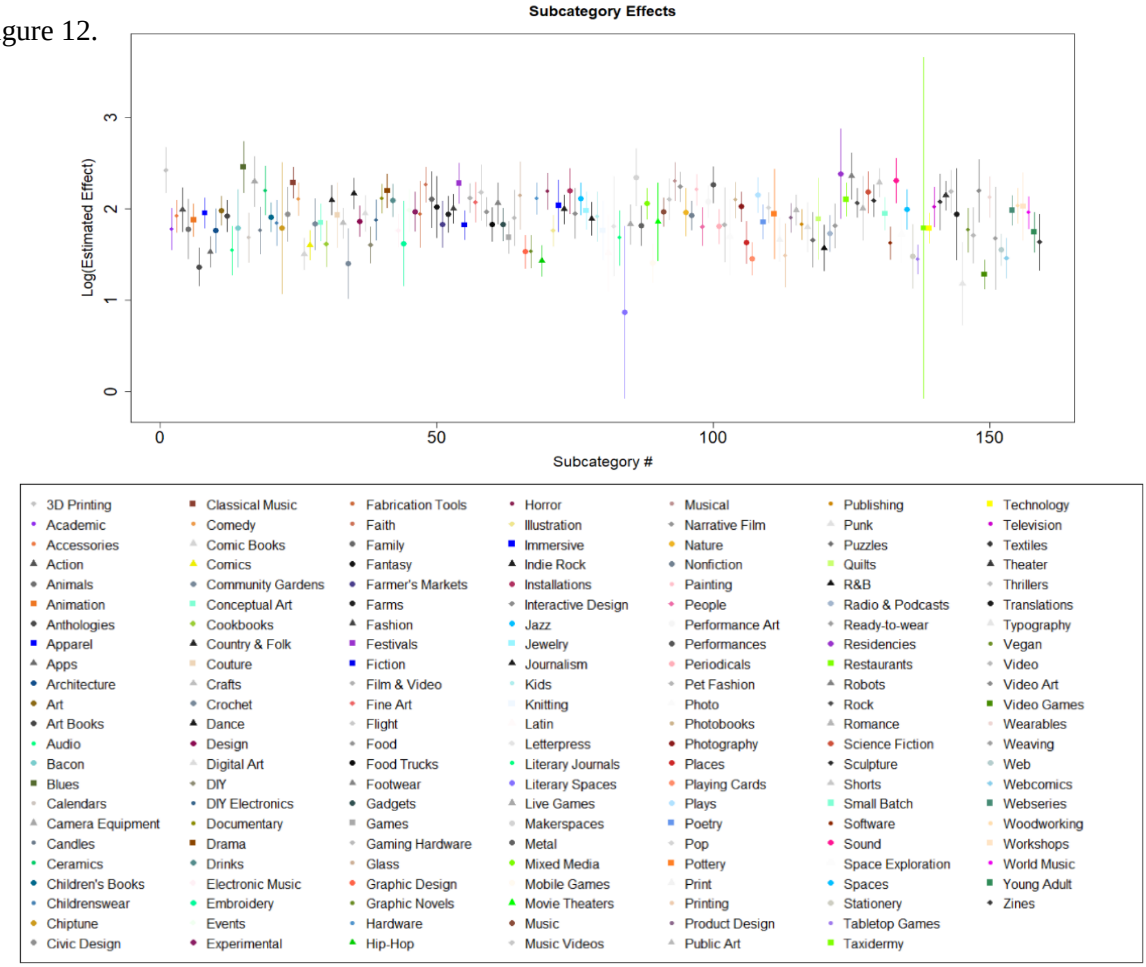


Figure 13 shows the diagnostic plots of our final model. From these plots, we can see that the spread of the residuals and the |standardized residuals|0.5 are minimal. However, we fail to have a constant mean function, constant variance, the standardized residuals are not approximately normal. This could due to the large association between backers and the subcategories. Removing backers would make the plots better, but would mask the true problem. Backers provide important information that cannot be ignored. Given more time and resource, it may be possible to reduce background noise and improve this model.

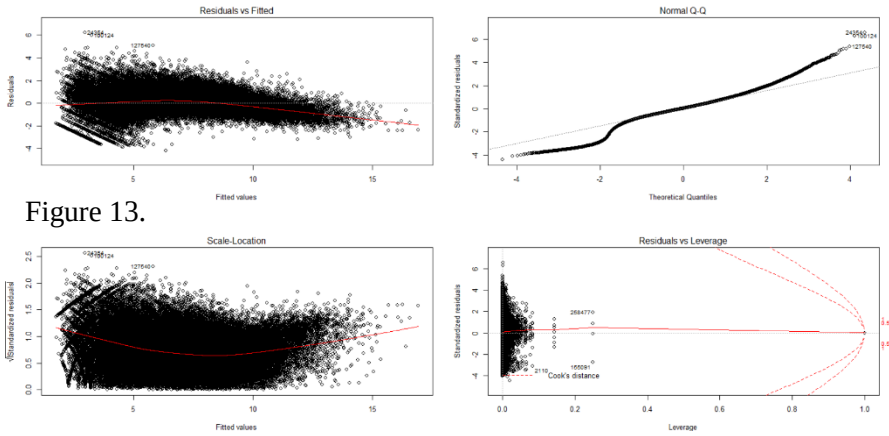
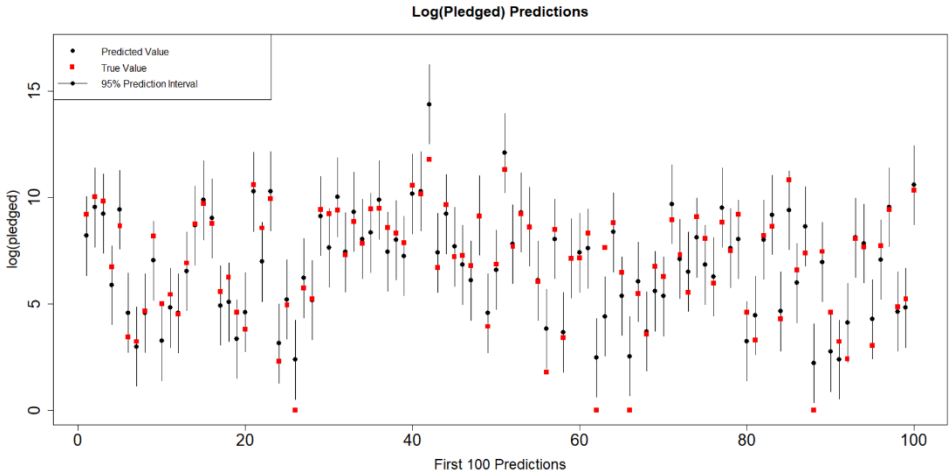


Figure 13.

To test this model, we made predictions using the test dataset. An MSE was calculated using the log(pledged) from the test data. The value of the MSE for this model is 0.893. The MSE is roughly equal to the variance of the differences between the true log(pledged) and the predicted log(pledged). Therefore, $\log \frac{pledged_T}{e^{pledged_{pre.}}}$ has an approximate standard deviation of 0.945. This ratio has a 95% statistical interval of (-1.89, 1.89) and the exponential of this ratio has a 95% statistical interval of (0.151, 6.619). Figure 14 shows the first 100 prediction on the test data in terms of log(pledged).

Figure 14.



Conclusions

Based on our logistic regression model, evidence supports our first hypothesis. However, based on our model, evidence does not support the second hypothesis that the runtime of a project has a negative effect on the success of a project. Finally, based on our model, evidence suggests that the type of project does have some effect on the success of project, supporting our fourth hypothesis. Each effect is various in size and whether it is positive or negative, so one must be very careful when choosing the type of Kickstarter to pursue.

Based on our linear regression model, contrary to our third hypothesis, goal has a positive effect on the amount pledged to a project. However, based on our model, evidence does support the fifth hypothesis that the type of project has an effect on the amount of money raised. As with the logistic model, the type of project has a varied effect, so one must be careful when choosing the type of Kickstarter to pursue.

It's also interesting to note that there is an interesting relationship between the launch month of a Kickstarter and its amount of money pledged. As seen in Figure 10, there was a peak in amount pledged for Kickstarters launched in May, with a smaller, but still significant rise leading up to October as well. In stark contrast, there is a dip in amount pledged during July, October, and December.

The reason behind this relationship might be explained by the demographics Kickstarter users. The majority of Kickstarter users have incomes between $0K and $50K and are college educated. In terms of age, 30% of Kickstarter users are 25-35 years old, with the next largest age

groups being 18-24 at 19% and 35-44 at 19%. With this in mind, it is possible that the dip in amount pledged in July, October, and December are due to the holidays. Considering the low income of the average Kickstarter user, it's possible that they need to be more frugal in these months due to spending around holidays such as the 4th of July, Halloween, and Christmas. However, confirming the link between demographics and Kickstarter variables would require further research.