Understanding Successful

Kickstarter Campaigns

Davis Busteed — ECON 388

## 

Table of Contents

[INTRODUCTION 1](#_Toc6241873)

[DATA 1](#_Toc6241874)

[Data Source 1](#_Toc6241875)

[Data Description 2](#_Toc6241876)

[MODEL ESTIMATION 4](#_Toc6241877)

[RESULTS 8](#_Toc6241878)

[Assumptions 8](#_Toc6241879)

[Significance Testing 9](#_Toc6241880)

[Interpretation of Estimators 10](#_Toc6241881)

[CONCLUSION 11](#_Toc6241882)

[Appendices 13](#_Toc6241883)

[Appendix A – Data Dictionary 13](#_Toc6241884)

[Appendix B – Descriptive Statistics 14](#_Toc6241885)

[Appendix C – Correlation Matrices 15](#_Toc6241886)

[Appendix D – Final Model Specification 16](#_Toc6241887)

## INTRODUCTION

Kickstarter is an online fundraising platform, where anyone can submit a project idea and ask for donations. Compared to other online fundraising platforms, Kickstarter doesn’t engage in equity crowdsourcing. Although it is common for project owners to promise “gifts” to the first 100 or so supporters, donations to a Kickstarter campaign are not investments. Despite this, Kickstarter has facilitated $4.2 billion in donations and 161,152 successful fundraising campaigns since its launch in 2009.[[1]](#footnote-1)

Just like traditional fundraisers, Kickstarter campaigns have a target amount to raise, known as the goal. Interestingly enough, it appears that around 39% of Kickstarter campaigns fail to meet their goal.[[2]](#footnote-2) This begs the question, why do some Kickstarter fundraisers succeed while others fail? What factors contribute to a campaign’s success? How do these factors individually contribute to the total amount raised?

## DATA

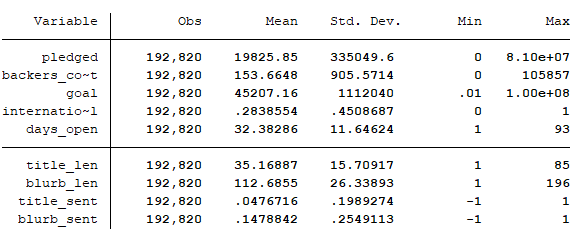
### Data Source

To understand what factors contribute to the amount pledged to a Kickstarter campaign, I looked for a dataset that contains the results of past fundraisers. While there were many different Kickstarter datasets available on Kaggle.com, I chose to use a dataset from WebRobots.io, an online service that publishes data from their automated web scrapers.[[3]](#footnote-3) The dataset included 50+ CSV files, each with around 4000 records. I used a Python script to combine these files, remove unnecessary columns, and perform some feature engineering. Records with empty fields were also removed during this point, ensuring that every record has a value for each variable. This script, as well as the cleaned dataset, can be found at <https://github.com/dbusteed/econometrics>.

### Data Description

After filtering out uncompleted fundraisers, I was left with 192,820 records outlining the details and results of each campaign. Some of the variables in the dataset were a product of the user’s “fundraiser settings.” When a user launches a fundraiser, they give their project a title and description, set a target goal amount, and choose the launch and end date of the campaign.

Other variables were simply recorded after the campaign had finished, such as the number of actual supporters a campaign had. Additional values were calculated during the initial data processing. For example, the variable *days\_open* was calculated as the difference between the deadline and the launch date. See Appendix A for a complete description of variables.

A summary of the main variables is as follows:

The processed dataset also includes two groups of dummy variables. During the data processing in which the dummy variables were created, one dummy from each group was dropped so that the variables wouldn’t cause perfect collinearity. By default, the Pandas data analysis module used in this step drops the first dummy variable in alphabetical order.

The first group of dummy variables was the category the fundraising project most closely aligns with. There are 15 total categories available on the Kickstarter platform, ranging from food to technology. As alphabetical order dictates, the art category was dropped from the set of dummies, and must be inferred. See Appendix 2.1 for a summary of this group.

The second group of dummy variables are the month that the campaign was launched. Similarly, the dummy for the month of April is not shown. See Appendix 2.2 for a summary of the launch month group.

## MODEL ESTIMATION

With the dataset cleaned and defined, I began specifying and estimating regression models. Unless stated otherwise, all regressions were estimated using the Ordinal Least Squares (OLS) method. If certain assumptions hold, the OLS estimator will be unbiased and efficient. I started with the following simple model. I thought that these RHS variables would provide a decent amount of explanatory power on the amount pledged to a campaign.

(1)

Originally, I experimented with both *pledged* and its log transformation as the dependent variable of this regression, but decided to use *log(pledged)* for future modeling. I wouldn’t be able to compare R2 from regressions with different LHS variables, and since I anticipated using *log(pledged)* in future modeling, it made sense to also use it in this first model.[[4]](#footnote-4) When estimated with OLS, equation (1) gave the following coefficients:

(2)

As I looked at the relatively small magnitudes for and in equation (2), I decided to apply natural log transformation to these terms. I started with *goal*, and got the following results upon estimation:

(3)

With still quite small, I attempted a log transformation on *backers\_count*. The results of this evaluation, seen below, were quite surprising:

(4)

By taking the natural log of *backers\_count*, the R2 increased by about 900%. A comprehensive interpretation of all coefficients will be given once the regression specification is complete, but until then, I continued to tweak the regression model. I removed *days\_open* and *international* from the model, and found no significant decrease in R2, which hinted at possible overspecification. Although this technically lowers the degrees of freedom, I don’t think including these irrelevant variables would cause too many issues since the data includes 192,820 observations. That said, I still chose to remove them from the regression.

During this step, I also experimented with the four calculated fields (length and sentiment scores for the title and blurbs of the project) to see if they added any explanatory value. After removing statistically insignificant variables (those with p-values under 0.01), I was left with the following specification:

(5)

To better explain the variation in *log(pledged)*, I estimated a regression that includes the dummy variables for project category. The specification for this model was the same as equation (5), but with following . I quickly learned, however, that this regression specification had perfect collinearity, violating one of the basic Gauss-Markov assumptions for unbiased estimates in OLS. By calculating a correlation matrix for all category dummy variables, I found that the dummy variable for photography and film had a perfect correlation of (see Appendix C.1). In order to use these dummy variables with confidence, I combined the photography and film categories into one variable. The resulting regression model was as follows:

(6)

Estimating this equation with OLS gave the following:

(7)

In addition to this, I experimented with another set of regression models with the month of launch dummy variables in the following form:

(8)

This regression had similar results, but some of the month dummy variables weren’t significant, so I decided to continue working with equation (6). I also tried regressing both sets of dummy variables (category and launch month) on *log(pledged)*, but found that it gave no additional insight. Although losing degrees of freedom wasn’t an issue since , it seemed to unnecessarily over-specify the model.

For this reason, it seemed that the regression model defined in equation (6) was the most appropriate for understanding the success of Kickstarter campaigns. Before continuing on, I checked to see whether this regression was homoscedastic or not. Since the number of variables in the regression specification would be too tedious for the White test, I performed a Breusch-Pagan test to test for homoskedasticity. I saved the residuals from the regression estimated in equation (7), then regressed the square of these residuals on the RHS variables, and estimated to following equation:

(9)

The overall significance of this estimation was given as , which is extremely significant. With proof that the regression model exhibits heteroskedasticity, I re-estimated the regression model from equation (6) using robust standard errors. The coefficients and R2 didn’t change from the estimation found in equation (7), instead, only the standard errors did. This in turn affected the test statistics, confidence intervals, and overall F-statistic, which will be addressed in the next section. See Appendix D for the full coefficients of the estimated model, as well as the difference between the normal and robust standard errors.

## RESULTS

### Assumptions

As mentioned earlier, certain assumptions must hold for OLS to be an unbiased estimator. First, the population model must be linear in the parameters. By looking at equation (6), it is clear that this assumption holds.

Next is random sampling. Data collection for this research was done with a web scraper. These automated programs often utilize randomness to limit the pages they visit while still collecting a random sample, therefore, I felt comfortable with this assumption.

The third assumption of perfect collinearity was address earlier, but can be mentioned once again. After combining the dummy variables for the film and photography categories, there appears to be no multicollinearity issues among the RHS variables (see Appendix C.2 for a full correlation matrix).

The fourth assumption of the Gauss-Markov Theorem (Zero Conditional Mean) is expressed as . Although this assumption can’t be tested, I don’t believe that all the RHS variables in the model are strictly exogenous. For example, I expect that factors such as the perceived possibility of a project, which can’t be measured, are included in the error term. Because it seems logical that users are willing to support a project if they believe that the it will actually come through, then , which also means that .

Since I had no available variables in the dataset to use as an instrument for *backers\_count*, I accepted that my estimates would be slightly biased. The error term in unobservable, so it is improbable to provide a detailed estimate of the bias’s magnitude. Despite this, I still inferred that the bias will be positive, which means that   is probably an overestimate of the true parameter. Since the 85% of the variation in *pledged* is already explained by the RHS variables, I expect that ’s bias will be minor.

As for the final Gauss-Markov assumption, it was already demonstrated that regression model demonstrates heteroskedasticity, which was controlled by using robust standard errors.

### Significance Testing

The final regression model was extremely statistically significant overall (F-statistic was 48913). For good measure, a joint significance test on the category dummy variables was conducted in the following manner:

The F-statistic of this hypothesis test is 683.47, which is well above any of the standard significance cutoffs. Even with robust standard errors, it is clear that each input variable is individually statistically significant. The lowest test statistic was 6.75, which in itself is quite significant.

### Interpretation of Estimators

The final estimated model, which can be seen in equation (6) or Appendix D, contains 17 estimators. Before interpreting the magnitude and sign of these estimators, it is important to remember that the original regressand underwent a log transformation, so that every model tested used *log(pledged)* as the dependent variable.

Starting with the intercept estimate of 1.03, this is the average value of *log(pledged)* when the other terms evaluate to zero. This will be the case when the fundraiser is in the art category. Since it is unlikely that everything else will evaluate to zero, this intercept isn’t very interesting; instead, it is included for technical reasons.

Next is and  , which have similar interpretations, being that they are both elasticities. is the elasticity of *pledged* with respect to *backers\_count*. As seen earlier in the model testing, including *backers\_count* with a log-log relationship increased the R2 significantly. For a 1% change in the number of backers a campaign has, the pledged amount is predicted to encounter a 1.55% increase. Although I know that is a bit biased, this seems like a reasonable elasticity for this relationship.

Similarly, although with less magnitude, is the elasticity of *pledged* with respect to *goal*. For a 1% change in the target goal of the fundraised, the predicted pledged amount increases by 0.083%. This slight positive elasticity is also quite reasonable, as bigger projects are more likely to get more funding, although the goal alone doesn’t dictate a fundraiser’s success.

In my opinion, the interpretation of is the most straightforward. For every character added to the title of a Kickstarter campaign, the predicted pledged amount raises by 0.42%. This may not seem like much, but, for example, when 10 more characters are added to the project title, it is predicted that amount pledged will increase by 4.2%. I find this completely logical, being that the title of one’s fundraiser is extremely important in gaining another user’s interest.

The estimators for the category dummy variables ( through ) have an interesting interpretation. As mentioned above, the constant represents the art category. Each then, represents the average percent change in *pledged* when the project is of the given category () compared to when the project is in the art category. For example, compared to art, when a fundraiser is comics, the predicted percent change on the total amount raised drops by 68%. On the other hand, if the category is theater, the predicted percent change is 20% higher than if it were a campaign for an art project. With this interpretation, it appears that the dance category yields the highest chance of funding, while the games category is the lowest.

Overall, the final model had an R2 of 0.8543, which suggests that the RHS variables in the regression specification explain the 85.43% of the variation in the total amount pledged.

## CONCLUSION

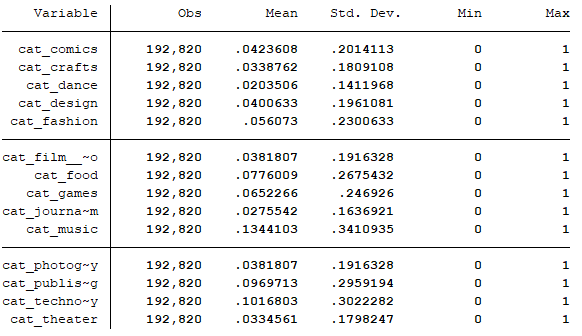
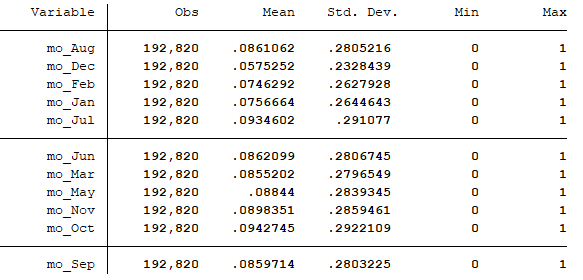
Although it is impossible to scientifically measure the ingenuity of the Kickstarter campaigns that are bound for success, it does appear that some quantitative and qualitative factors can be linked to the amount a fundraiser actually raises. It seems that a campaign’s pledged amount will increase as an appropriate goal, category, and descriptive title are selected. Most importantly, however, is the number of backers a campaign has. Even if each individual contribution is small, the data suggests that increasing the number of backers is key to predicting the success of a Kickstarter campaign.

## Appendices

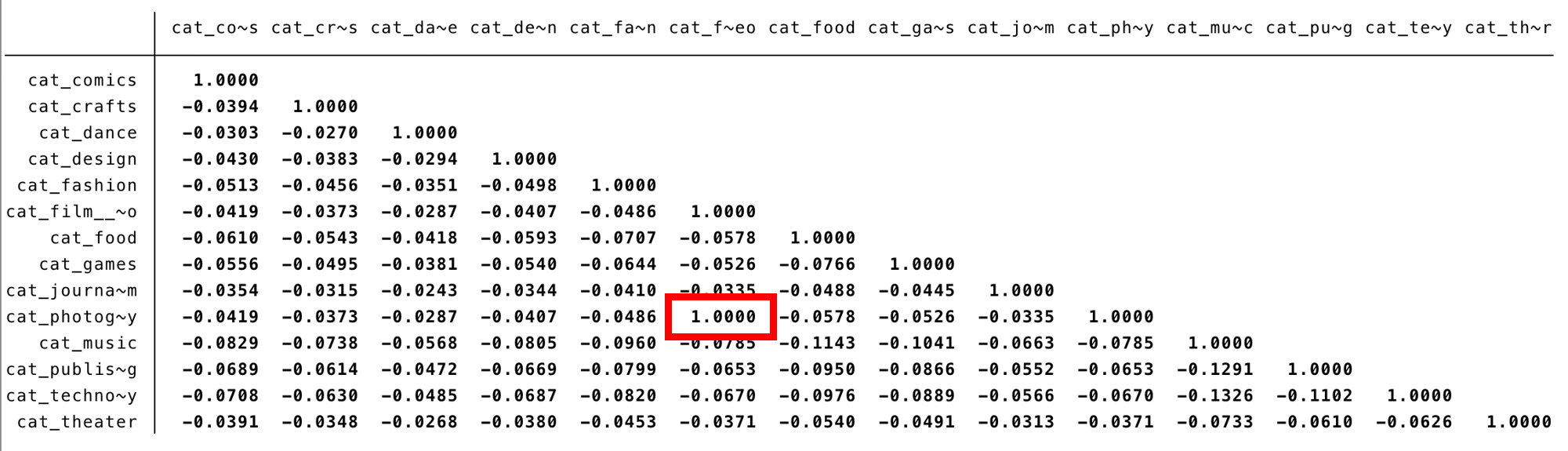
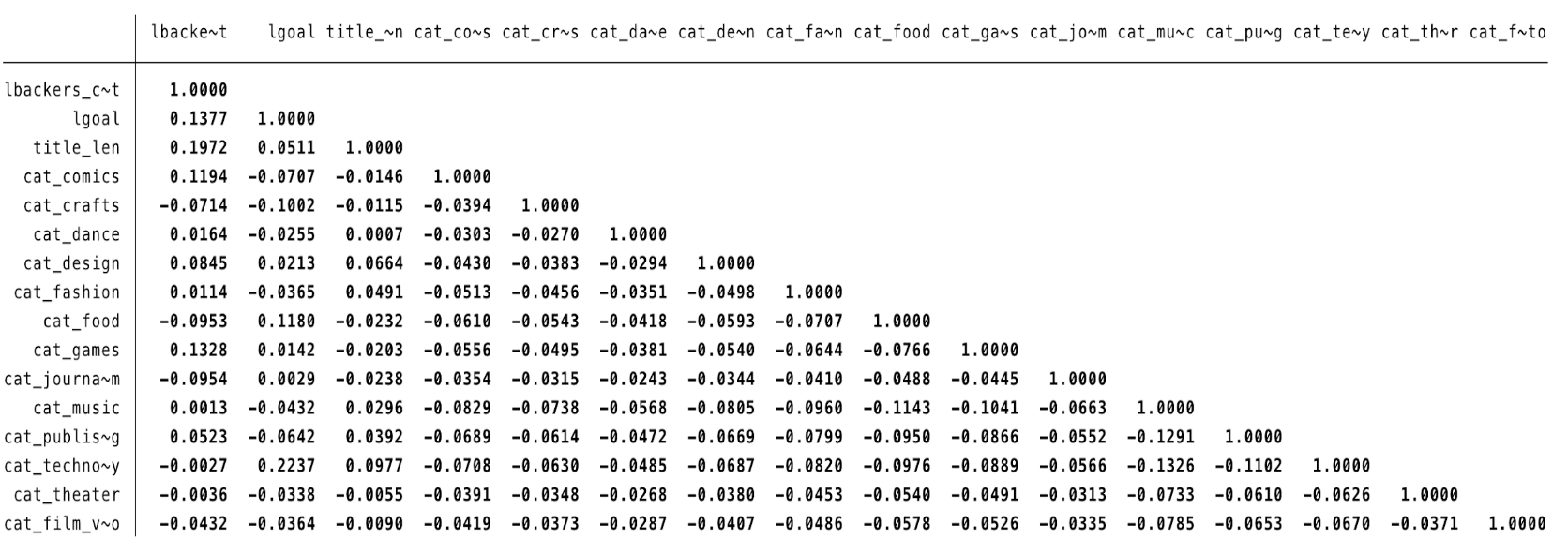
### Appendix A – Data Dictionary

|  |  |  |
| --- | --- | --- |
| Name | Description | Notes |
| pledged | The total amount pledged or donated to the campaign | This is the dependent variable of interest |
| backers\_count | The number of users who supported a campaign by pledging money |  |
| goal | The target fundraising amount. This is set by the project owner when launching the campaign | A fundraiser is successful if the pledged amount exceeds the goal |
| days\_open | The number of days the campaign was open (when other users were able to donate) | This was calculated by taking the difference of the ‘deadline’ and ‘launched\_at’ in the raw dataset |
| international | A dummy variable for whether the campaign was international or not | 0 = domestic (US)  1 = international |
| title\_len | Number of characters in the campaign’s title | Calculated with the “len” method in Python |
| blurb\_len | Number of characters in the campaign’s description, also known as the blurb | Same as above |
| title\_sent | Sentiment score of the campaign’s title. | Sentiment scores range from -1 (negative) to 1 (positive) |
| blurb\_sent | Sentiment score of the campaign’s blurb. | Both sentiment scores calculated with the TextBlob module in Python |
| cat\_\* | Dummy variables for the fundraiser category | 0 = not of that category  1 = of that category |
| mo\_\* | Dummy variables for the month the campaign was launched | 0 = if not that month  1 = if that month |

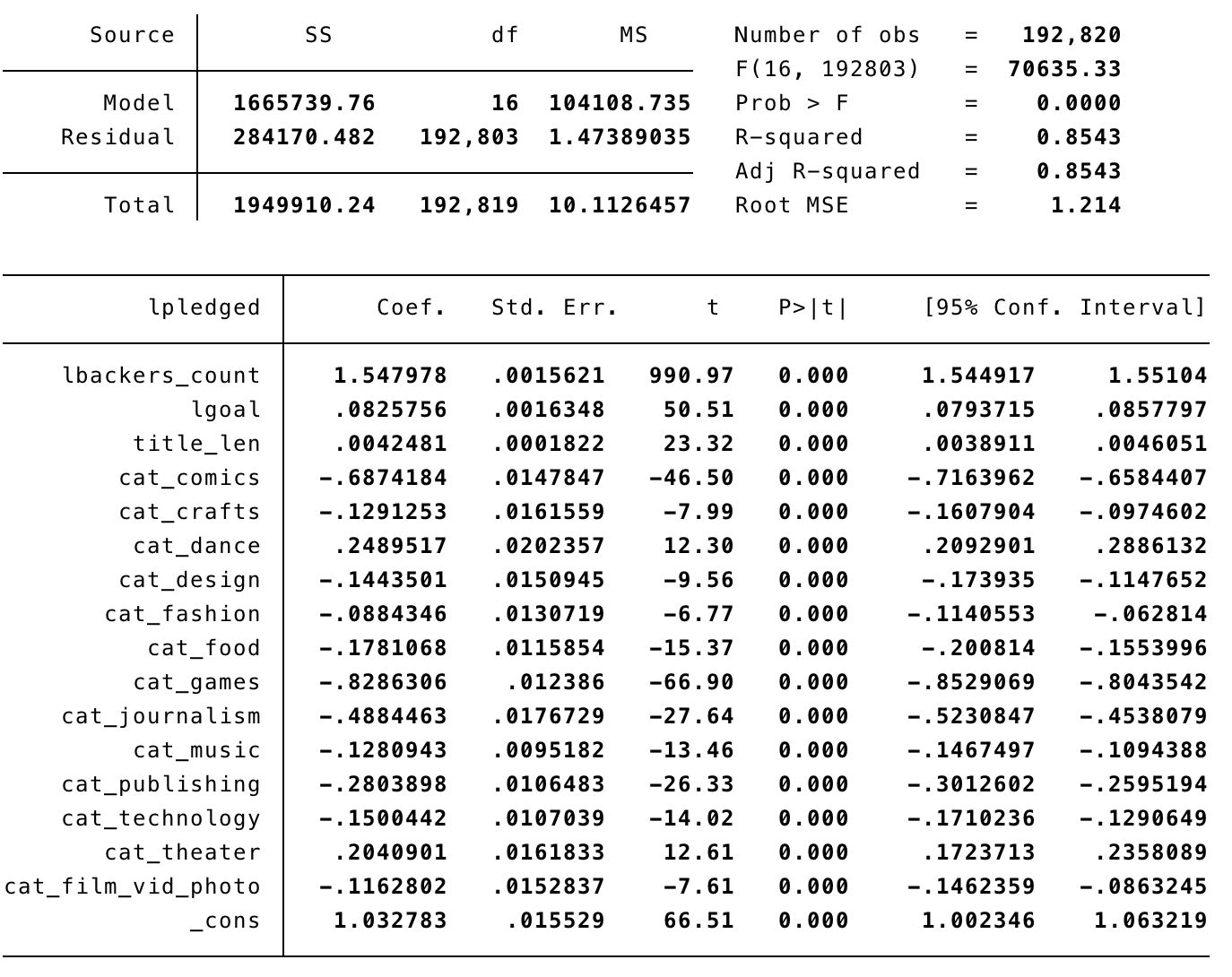
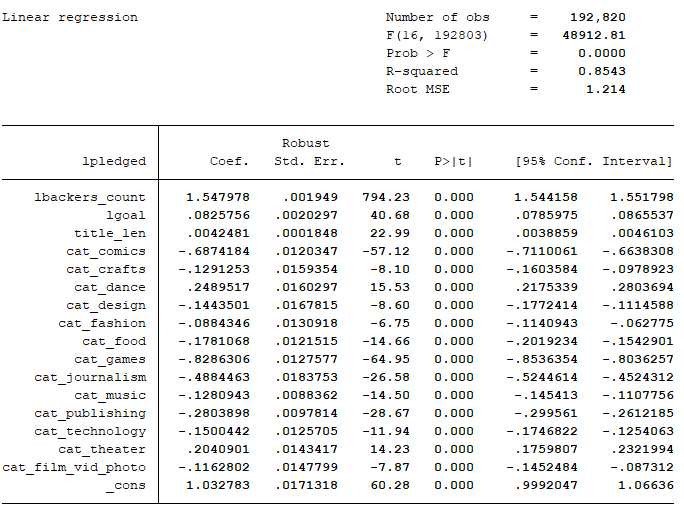
### Appendix B – Descriptive Statistics

1. Summary of category dummy variables (cat\_art dropped) 
2. Summary of launch month dummy variables (mo\_Apr dropped)

### Appendix C – Correlation Matrices

1. Correlation matrix of all categories
2. Correlation matrix of all RHS vars from equation (6)

### Appendix D – Final Model Specification

1. With normal standard errors
2. With robust standard errors

1. https://www.kickstarter.com/about [↑](#footnote-ref-1)
2. Calculated from the cleaned data set [↑](#footnote-ref-2)
3. https://webrobots.io/kickstarter-datasets/ [↑](#footnote-ref-3)
4. When transforming *pledged* and *backers\_count*, the following Stata command was used to avoid taking the natural log of zero:  gen logx = log(x+1). [↑](#footnote-ref-4)