Lab 2 Research Proposal

Oleg Ananyev, Oren Carmeli, Romain Hardy, Sam Rosenberg

Fall 2021

Introduction

> Lab 2 Intructions: Your introduction should present a research question and explain the concept that you're attempting to measure and how it will be operationalized. This section should pave the way for the body of the report, preparing the reader to understand why the models are constructed the way that they are. It is not enough to simply say, "We are looking for product features that enhance product success." Your introduction must do work for you, focusing the reader on a specific measurement goal, making them care about it, and propelling the narrative forward. This is also a good time to put your work into context, discuss cross-cutting issues, and assess the overall appropriateness of the data.

The Google Play Store is one of the major hubs for applications for Android mobile phone and tablet users. Users are can download any app on the Google Play Store for personal consumption, however in a sea of tens of thousands of apps or more, we would like to determine the drivers for app downloads. Given that most phones have limited screen real estate and they are intended for light computational tasks, getting an app in-front of a consumer is a challenging proposition as developers have several factors going against them (e.g., vast quantity of duplicate apps, limited screen real estate for ads, limited screen time by most users, etc.).

Statista reports there were ~220 billion mobile application downloads globally in 2020. Another study found that ~70% of all US digital media is spent on mobile applications. Mobile app usage is becoming the focal interaction medium in which organizations can engage and help their customers and it is useful for organizations to further understand drivers that help lead to more mobile application usage and a better customer experience. This dataset allows to analyze the potential levers to app downloads such as average review rating, price, app size, app category, etc. With this dataset, we will plan to build a statistical linear model to determine the importance (or unimportance) of each of these variables on app downloads. Given our beliefs on what causes individuals to download a mobile application we believe there are some important ommitted variables that can influence app downloads such as brand awareness from marketing campaigns, the size of existing customer bases, seasonality, and individual life circumstances. Despite these limitations, we believe the dataset provides useful information to evaluate the importance of other characterstics outside of these external factors.

Data and Research Design

> Lab 2 Intructions: After you have presented the introduction and the concepts that are under investigation, what data are you going to use to answer the questions? What type of research design are you using? What type of models are you going to estimate, and what goals do you have for these models?
> **Research Question: Does average rating increase the number of downloads a Google Play mobile application accumulates. This is a causal question.

To investigate our research question, we will leverage a publicly available data set of apps available on the Google App Store with key information such as downloads, file size, rating, category, price, etc. This data set was scrapped in 2019 and made available on Kaggle.com for use by anyone with interest. It contains records of ~10,000 app available on the Google Play Store. Given the size of our dataset, we will use OLS regression to answer our research question. OLS regression is a plug-in estimator of the best linear predictor for the joint distribution of various random variables. In this case, the random variables of interest include the various attributes of Google Play mobile applications. The specific variables we will use include:

Dependent Variable:

- installs = total installs the application accumulated

Independent Variables:

- size = the memory space the application takes up
- reviews = total reviews the application accumulated. This won't be an independent variable, it will be used to filter out outliers. More details in our EDA section
- rating = the average rating for the application (out of 5 stars).
- price = the price to download the application (0 = free)
- category = the category tagged for the application (i.e Lifestyle)
- content_rating = rating assigned to the application (i.e Teen, Everyone, Mature 17+)
- type = Options include Free (free to download) or Paid (costs money to download)

OLS regression is built with a goal of making predictions on new data. But, we see believe it will help answer the causal question describing the relationships across our independent and dependent variables. We believe that through exploratory data analysis, measuring the prediction power of the model, and conceptualizing the coefficient estimates we can properly inform and develop a statistically sound argument around our research question.

Model Building

> Lab 2 Intructions: You will next build a set of models to investigate your research question, documenting your decisions. Here are some things to keep in mind during your model building process: 1. *What do you want to measure*? Make sure you identify one, or a few, variables that will allow you to derive conclusions relevant to your research question, and include those variables in all model specifications. How are the variables that you will be modeling distributed? Provide enough context and information about your data for your audience to understand whatever model results you will eventually present. 2. What covariates help you achieve your modeling goals? Are there problematic covariates? either due to *collinearity*, or because they will absorb some of a causal effect you want to measure? 3. What *transformations*, if any, should you apply to each variable? These transformations might reveal linearities in the data, make our results relevant, or help us meet model assumptions. 4. Are your choices supported by exploratory data analysis (*EDA*)? You will likely start with some general EDA to *detect anomalies* (missing values, top-coded variables, etc.). From then on, your EDA should be interspersed with your model building. Use visual tools to *guide* your decisions. You can also leverage statistical *tests* to help assess whether variables, or groups of variables, are improving model fit. At the same time, it is important to remember that you are not trying to create one perfect model. You will create several specifications, giving the reader a sense of how robust (or sensitive) your results are to modeling choices, and to show that you're not just cherry-picking the specification that leads to the largest effects. At a minimum, you need to estimate at least three model specifications: The first model you include should include *only the key variables* you want to measure. These variables might be transformed, as determined by your EDA, but the model should include the absolute minimum number of covariates (usually zero or one covariate that is so crucial it would be unreasonable to omit it). Additional models should each be defensible, and should continue to tell the story of how product features contribute to product success. This might mean including additional right-hand side features to remove omitted variable bias identified by your casual theory; or, instead, it might mean estimating a model that examines a related concept of success, or a model that investigates a heterogeneous effect. These models, and your modeling process should be defensible, incremental, and clearly explained at all points. Your goal is to choose models that

encircle the space of reasonable modeling choices, and to give an overall understanding of how these choices impact results.

Based on our exploratory data analysis we decided on various interaction terms and transformations to use in our linear model. Our focus was to maximize its prediction accuracy (Rˆ2) as well as maintaining a model that is explanable.

Model Limitations

5a. Statistical limitations of your model

> Lab 2 Intructions: As a team, evaluate all of the large sample model assumptions. However, you do not necessarily want to discuss every assumption in your report. Instead, highlight any assumption that might pose significant problems for your analysis. For any violations that you identify, describe the statistical consequences. If you are able to identify any strategies to mitigate the consequences, explain these strategies. Note that you may need to change your model specifications in response to violations of the large sample model.

To validate that OLS is the most appropriate method to analyze our data, we will validate all 5 assumptions; I.I.D, no perfect collinearity, linear conditional expectation, homoskedastic errors, normally distributed errors.

- **I.I.D**: According to the Kaggle authors, this data set was collected by randomly scraping the Google Play Store. Since no clusters of applications were specifically targeted, we can reasonably use the entirety of the store as our reference population. We recognize that applications likely have some degree of interdependence, especially within genres. For example, the success of one application probably has a negative impact on other applications of the same type. Due to the large size of this data set, however, we expect any dependencies to be negligible. We also have reason to believe that the data are identically distributed, as they are drawn from the same population of applications. One could argue that since the Google Play Store changes over time, the distribution also shifts in response. Because the authors do not mention the time frame across which the data was collected, we will assume that they originated from a single snapshot of the Play Store and that no shifts in the underlying distribution occurred.
- **No Perfect Collinearity**: We can immediately conclude that `log_installs`, `log_reviews`, `rating`, and `log_size` are not perfectly colinear as otherwise the regression above would have failed. We can also assess near perfect colinearity for these variables by observing the robust standard errors returned by the regression model. In general, highly colinear features will have large standard errors. Since the standard error of the coefficients are small relative to their magnitude, we can reasonably conclude that they are not nearly colinear.
- **Linear Conditional Expectation**: To verify the assumption of linear conditional expectations, we seek to show that there is no relationship between the model residuals and any of the predictor variables. That is, the model does not systematically underpredict or overpredict in certain regions of the input space. Plots 1 through 3 show the relationships between the model residuals and individual predictors. The residuals are generally well-centered around zero, although the model seems to underpredict when `log_reviews` is high and `rating` is low. The fourth plot shows the model residuals as a function of the model predictions. Here, the model seems to underpredict in the left-most and right-most regions, and slightly overpredict in the middle. Overall, there are no strong non-linear relationships between the model residuals and the input features, and we do not find enough evidence to reject the assumption of linear conditional expectation.
- **Homoskedastic Errors**: When assessing homoskedastic errors, we seek to determine if there is a relationship between the variance of the model residuals and the predictors. If the homoskedastic assumption is satisfied, then we should observe a lack of relationship; conversely, if the data are heteroskedastic then the conditional variance will depend on the predictors. The first plot is an eyeball test of homoskedasticity, showing the model residuals as a function of the model predictions. We notice that the spread of the residuals is mostly consistent throughout the data, although the right-hand side is somewhat narrower. As a more concrete assessment, we also perform a Breush-Pagan test with the null hypothesis that there are no heteroskedastic errors in the model. Since the p-value falls below our

significance threshold of 0.001, we find enough evidence to reject the null hypothesis. In response to this failed assumption, we report robust standard errors (adjusted for heteroskedasticity) instead of non-adjusted errors.

- **Normally Distributed Errors**: When assessing the normality of the error distribution, we seek to determine if the model residuals are approximately Gaussian. If so, then the sample quantiles of the residuals should closely match the theoretical quantiles of a normal distribution in a Q-Q plot. Below, we plot the Q-Q plot associated with our model. In general, the residuals seem to follow a normal distribution, as the middle quantiles match the corresponding theoretical quantiles. However, the tails of the residual distribution are fatter than expected; the first quantiles occur at smaller than expected values, and the last quantiles occur at larger than expected values. Overall, the assumption of normally distributed errors seems imperfect but reasonably justified.

5b. Structural limitations of your model

Lab 2 Intructions: What are the most important *omitted variables* that you were not able to measure and include in your analysis? For each variable you name, you should *reason about the direction of bias* caused by omitting this variable and whether the omission of this variable calls into question the core results you are reporting. What data could you collect that would resolve any omitted variables bias?

Results

Lab 2 Intructions: You should display all of your model specifications in a regression table, using a package like `stargazer` to format your output. It should be easy for the reader to find the coefficients that represent key effects near the top of the regression table, and scan horizontally to see how they change from specification to specification. Make sure that you display the most appropriate standard errors in your table. In your text, comment on both *statistical significance and practical significance*. You may want to include statistical tests besides the standard t-tests for regression coefficients. Here, it is important that you make clear to your audience the practical significance of any model results. How should the product change as a result of what you have discovered? Are there limits to how much change you are proposing? What are the most important results that you have discovered, and what are the least important?

Conclusion

Lab 2 Intructions: Make sure that you end your report with a discussion that distills key insights from your estimates and addresses your research question

Appendix