Lab 2 Research Proposal

Oleg Ananyev, Oren Carmeli, Romain Hardy, Sam Rosenberg

Fall 2021

Contents

1 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.

In 2020, Statista reported a total of ~220 billion mobile application downloads globally. Another study found that ~70% of all US digital media is spent on mobile applications. As mobile applications become one of the the principal media through which organizations engage with their customers, it becomes increasingly valuable to understand the drivers that lead to greater usage and better customer experience.

The Google Play Store is one of the major hubs for Android mobile phone and tablet applications. Users can download any application on the Google Play Store for personal consumption across a wide range of categories. Making an application stand out in a sea of thousands of competing applications is no trivial task, however. To succeed, developers must carefully consider factors such as price, application size, and genre. Another variable that may be critical to an application's success is consumer rating. Today, smart algorithms play a key role in suggesting applications to consumers, creating a feedback loop that propels certain applications towards success and leaves many others behind. An understanding of the relationship between consumer rating and application success would be incredibly valuable to developers seeking to create the next viral application.

The following study a causal analysis of the relationship between average consumer rating and application downloads. Using a data set scraped directly from the Google Play Store, we will build a linear model that assesses the importance (or lack of importance) of average consumer rating on the number of downloads, with additional variables such as price, application size, and application category serving as controls. Confirming the existence of a causal pathway would signal application developers to invest heavily in improving their ratings, for instance by buying positive reviews from consumers. Our study may also be of interest to Google, who has a vested interest in preventing developers from unfairly inflating their ratings.

Given our prior beliefs on what factors motivate individuals to download applications, we believe there are omitted variables not included in our data set that may influence application downloads. These include brand awareness from marketing campaigns, differences in existing customer bases, seasonality, and individual life circumstances, among possible others. In spite of these limitations, our study should provide useful insight into the factors that cause an application's success.

The paper is structured as follows. Section 2 outlines our research question, the causal model we will use to contextualize our regression analysis, and the research design. Section 3 describes our data, the explanatory variables we will include in the modeling phase, and the transformations we will apply. Section 4 contains our statistical models, and Section 5 discusses their significance as well as their statistical validity. Finally, in Section 6 we present our conclusions and discuss the implications of our results.

2 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?

The goal of this study is to assess the causal factors of application success. Specifically, we seek to determine if there is a statistically significant relationship between average consumer rating and application downloads. In the ensuing sections, we will answer the following question:

> *Does a higher average consumer rating lead to greater downloads for Google Play Store applications?*

In order to answer this question, we will leverage publicly available data about applications available on the Google Play Store. This data was randomly scraped from the Google Play Store interface and uploaded to Kaggle.com in 2019. It contains key information about each sampled application, such as downloads, file size, rating, category, and price. In total, the data contains records of ~10,000 applications.

For the modeling step, we will use OLS regression. OLS regression is the plug-in estimator for the best linear predictor of a dependent random variable given the joint distribution of a set of independent random variables. In our case, the dependent variable is our conceptualization of application success (the number of downloads an application receives), and the independent variables are the various attributes of Google Play Store applications. Although OLS regression is often used with the goal of making predictions on new data, here it will be used it to answer a causal question about the relationship between variables. By interpreting model coefficients within the context of a causal theory, we will develop a statistically valid argument that answers the research question.

Our analysis will make use the following features:

Dependent Variable:

- `installs` — The accumulated number of installs since the application was uploaded to the Google Play Store.

Independent Variables:

- `size` — The memory space occupied by the application.
- `reviews` — The accumulated number of reviews since the application was uploaded to the Google Play Store. This feature will be used to remove outliers (see the EDA section).

- `rating` — The average consumer rating for the application out of 5.
- `price` — The price of the application.
- `category` — The category tagged for the application (i.e Lifestyle, Game)
- `content_rating` — The official content rating given to the application (i.e Teen, Everyone, Mature 17+)
- `type` — The price type of the application. Possible options are Free (free to download) or Paid (costs money to download).
- `last_updated` — The date when the application was last updated.
- `current_ver` — The current version number of the application.

2.1 Causal Model

In this section, we describe the causal model which will serve as the reference point for our analysis. We also discuss possible omitted variables and the impact they may have on our study.

2.1.1 Simple Model

We identify five factors that bear causal influence on the success of an application. These are (1) consumer rating, (2) production quality, (3) category, (4) age, and (5) size. Our proposed causal diagram is shown below.

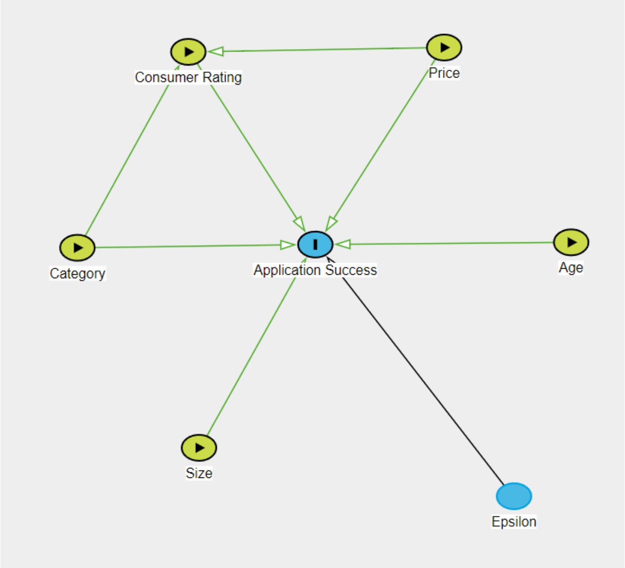

Figure 1: A hypothetical causal diagram for Google Play Store applications.

Application success

The goal of this analysis to identify causal factors for application success, and more specifically to assess whether consumer rating has a positive effect on application success. There are different ways in which we could operationalize application success. Our study will use the raw download count as a surrogate, since it directly measures the number of consumers that made the decision to download the application. Although this choice ignores other potential aspects of success such as revenue and social impact, it is effective in its simplicity and appropriate for an exploratory study.

Consumer Rating

Our independent variable of interest is consumer rating, which we are operationalizing using the `rating` field in the data set. We hypothesize that higher application ratings should cause greater downloads. This is because an application that has been highly rated is one that has been deemed worthwhile by other users.

When new consumers decide whether or not to download an application, they will likely trust the opinions of their peers and download applications with positive reviews. Conversely, we expect applications with negative ratings to have less success, since other users have judged them poorly. There should not be causal pathways from rating to any of the other explanatory variables, since they are determined during the development phase of the application while consumer rating is decided once the application is published to the Google Play Store.

Although we have not included it in the diagram, there is the possibility that a reverse causal path exists from application success to consumer rating. Successful applications are those that are enjoyable to a large number of consumers. Therefore, it is possible that users will rate applications with high download counts higher than applications with low download counts because they are primed to believe that such applications are better—otherwise, the successful applications would not have received so many views. Generally, we expect the reverse path to be weaker than the forward path. If ratings and downloads both have positive effects on one another, however, our models may suffer from positive feedback. We will discuss the implications of this effect in the results section.

Price

We expect price (`price` and `type` in the data set) to also have a causal effect on application success. Overall, we believe that paid applications should be more successful than free applications since they have more appealing features. This may not always be the case, however. For instance, free applications could have equally impressive features as paid applications, while paid applications could be lazily trying to cash out on naive users. We recognize that this assumption is imperfect, and we will discuss possible interpretations later in our analysis.

We also anticipate a causal pathway from price to rating, for similar reasons as the pathway from price to success. Specifically, we believe that consumers are likely to rate paid applications positively since those applications have more desirable features, whereas free applications will be reviewed negatively. Again, this causal pathway could lean in the opposite direction if our assumption about the relationship between price and application quality proves to be incorrect.

Category

The category that an application belongs to is likely to affect its success. Some types of applications like games and utilities appeal to a broad audiences and are therefore more likely to find success than applications which appeal to a smaller subset of consumers. This interaction may not necessarily be so straightforward, however. If an application belongs to a popular category, then it also has to compete with other applications in the same category, which may actually be detrimental to the application. Meanwhile, applications belonging to niche categories could have a greater chance of achieving success simply due to the fact that they have fewer competitors. Globally, we expect that the most successful applications to belong to popular categories, but that moderately successful applications can be found across almost a majority of categories.

Application category is probably also a predictor of consumer rating. Due to stylistic and functional differences between categories, it is likely that they are reviewed against different criteria. For example, a consumer reviewing a mobile game may place emphasis on the graphics, the fluidity of the controls, and the balance of the game mechanics, among others. A lifestyle application, on the other hand, will probably be judged on completely different features, such as ease of use, relevance in every day life, and usefulness. Because the review criteria depend on the category of the application, the ratings assuredly do too. It is difficult to predict in advance what categories are positively or negatively associated with consumer ratings, but we expect categories with narrower consumer bases to receive harsher ratings. Categories that can make users feel frustrated, such as games or dating applications, may also receive more negative reviews on average.

Age

If we are to interpret application success in terms of the number of downloads an application accumulates, then the age of an application is necessarily an influential factor. Applications can only receive more downloads as time passes, so the longer an application remains on the Google Play Store, the more downloads it can receive. As an example of why this is important, consider two applications: one that was uploaded one year ago, and one that was uploaded one week ago. The older application receives 100 downloads a month for the whole year while the newer application receives 1000 downloads in one week. If we were to only compare the download counts (1200 to 1000), it would seem as if the older application was more successful. By bringing age into the analysis, we are able to compare the applications by download rate instead of count, thus realizing that the newer application is far more successful than its older counterpart.

Unfortunately, our data set does not include an exact measure for age. However, the fields `current_ver` (the current version of the application) and `last_updated` (the date when the application was last updated) can serve as proxies for age. For instance, `current_ver` will generally be larger for older applications and smaller for newer applications.

We do not expect age to have causal pathways to any of the other explanatory variables, though it might be possible to argue that price and rating are affected by age. Unless our analysis clearly demonstrates otherwise, we will assume these effects are negligible.

Size

We foresee two opposing sides to the relationship between application size (represented by `size` in the data set) and application success. The first is a negative effect; given the limited space available on mobile and tablet devices, users may be more likely to download smaller applications. Alternatively, application size could be an indicator of production quality, in which case users may prefer larger applications over smaller ones. We hypothesize that the first effect takes precedence.

2.1.2 Omitted Variables

The true causal diagram is undoubtedly more complex than the one we have outlined above. We have identified several omitted variables that could affect our statistical models and any conclusions we infer from them, shown in the second causal diagram below. We discuss the relationships these omitted variables have on our existing variables and the ways in which they could bias our results.

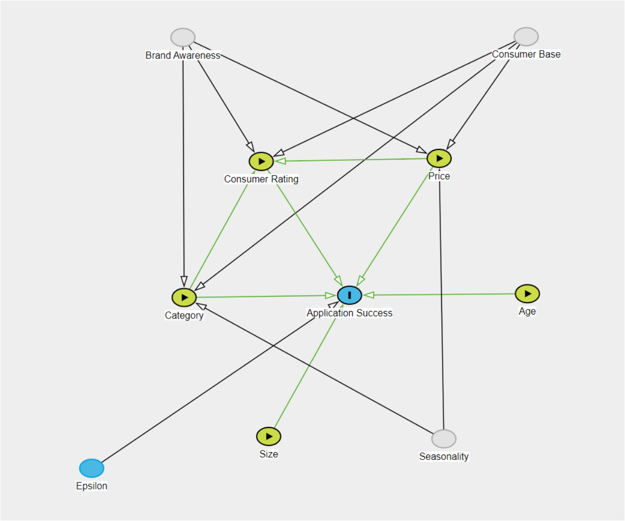

Figure 2: A revised version of the causal diagram including omitted variables.

Brand Awareness

Differences in Consumer Bases

Seasonality

Individual Life Circumstances

2.2 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 (R^2) as well as maintaining a model that is explainable.

```r
#install.packages('GGally')
#install.packages('moments')
#install.packages("corrplot")

install.packages('GGally', repos='https://ftp.osuosl.org/pub/cran/')
install.packages('moments', repos='https://ftp.osuosl.org/pub/cran/')
install.packages("corrplot", repos = "http://cran.us.r-project.org")
```

```r
library(GGally)
library(ggplot2)
library(lmtest)
library(lubridate)
library(moments)
library(sandwich)
library(stargazer)
library(tidyverse)
library(data.table)
```

```r
source('./functions/get_robust_se.R')
source('./functions/get_clean_dataset.R')

knitr::opts_chunk$set(echo = TRUE)
```

```r
data_clean <- get_clean_dataset()
glimpse(data_clean)
```

```
## Rows: 5,408
## Columns: 24
## $ installs            <dbl> 500, 500, 500, 1000, 1000, 1000, 1000, 1000, 1000,~
## $ size                <dbl> 22.0, 10.0, 2.5, 5.4, 2.4, 13.0, 6.5, 53.0, 15.0, ~
## $ reviews             <dbl> 156, 120, 124, 726, 112, 112, 123, 787, 141, 171, ~
## $ rating              <dbl> 4.6, 4.5, 4.8, 4.8, 4.5, 4.4, 4.8, 4.8, 5.0, 4.8, ~
## $ price               <dbl> 3.49, 0.00, 0.00, 3.99, 7.99, 0.00, 0.00, 1.99, 0.~
## $ is_free             <lgl> FALSE, TRUE, TRUE, FALSE, FALSE, TRUE, TRUE, FALSE~
## $ last_updated        <dbl> 0.402739726, 0.010958904, 1.958904110, 0.632876712~
## $ android_version     <dbl> 4.4, 5.0, 4.2, 4.1, 4.0, 4.4, 3.0, 4.0, 4.1, 2.2, ~
## $ current_version     <dbl> 1.1, 1.5, 1.2, 2.3, 3.0, 4.5, 3.1, 1.4, 1.8, 1.7, ~
## $ category            <chr> "MEDICAL", "DATING", "FAMILY", "BUSINESS", "PRODUC~
## $ is_family_category  <lgl> FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FA~
## $ is_game_category    <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, F~
## $ is_tools_category   <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, F~
## $ genre               <chr> "Medical", "Dating", "Entertainment", "Business", ~
## $ content_rating      <chr> "Everyone", "Mature 17+", "Everyone", "Everyone", ~
## $ is_content_everyone <lgl> TRUE, FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, T~
## $ type                <chr> "Paid", "Free", "Free", "Paid", "Paid", "Free", "F~
## $ install_group       <int> 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,~
## $ log_installs        <dbl> 2.69897, 2.69897, 2.69897, 3.00000, 3.00000, 3.000~
## $ log_size            <dbl> 1.3424227, 1.0000000, 0.3979400, 0.7323938, 0.3802~
## $ log_price           <dbl> 0.6522463, 0.0000000, 0.0000000, 0.6981005, 0.9537~
## $ log_current_version <dbl> 0.3222193, 0.3979400, 0.3424227, 0.5185139, 0.6020~
## $ log_last_updated    <dbl> 0.146977097, 0.004733502, 0.471130891, 0.212953395~
## $ log_reviews         <dbl> 2.193125, 2.079181, 2.093422, 2.860937, 2.049218, ~
```

```r
model_small  <- lm(log_installs ~ 1 + rating, data = data_clean)
model_medium <- lm(log_installs ~ 1 + rating + log_size + log_current_version +
                     log_last_updated + is_free + is_family_category +
                     is_game_category + is_tools_category + is_content_everyone,
                   data = data_clean)
model_large  <- lm(log_installs ~ 1 + rating + log_size + log_current_version +
                     log_last_updated + is_free + is_content_everyone +
                     rating * is_family_category + rating * is_game_category +
                     rating * is_tools_category,
                   data = data_clean)

# plot(model_small)
# plot(model_medium)
# plot(model_large)
```

```r
stargazer(
  model_small,
  model_medium,
  model_large,
  header = FALSE,
  type = "latex",
  se = list(get_robust_se(model_small), get_robust_se(model_medium),
            get_robust_se(model_large)),
  column.sep.width = "3pt",
  font.size = "small"
)
```

2.3 Model Limitations

2.3.0.1 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.

In the following section, we assess the five assumptions of the classic linear model: independence and identical distributions (I.I.D.), no perfect collinearity, linear conditional expectations, homoskedastic errors, and 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 entire set of applications on the Google Play Store as our reference population. We recognize that applications likely have some degree of interdependence, especially within categories. For example, the success of one application likely 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 make no specific mention of the time frame across which the data was collected, we will assume that they originate from a cross-sectional snapshot of the Google Play Store and that no shifts in the underlying distribution occurred during the sampling process.

No Perfect Collinearity

We can immediately conclude that the variables included in our models are not perfectly collinear, as otherwise the regressions above would have failed. We can also assess near perfect collinearity for these variables by observing the robust standard errors returned by the regression model. In general, highly collinear features will have large standard errors. Since the standard errors of the coefficients are small relative to their magnitude, we can reasonably conclude that they are not nearly collinear.

Linear Conditional Expectations

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. Figures show the relationships between the model residuals and metric-scale 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, so we do not find enough evidence to reject the assumption of linear conditional expectation.

Homoskedastic Errors

Table 1:

	Dependent variable:		
	log_installs		
	(1)	(2)	(3)
rating	0.448***	0.338***	0.288***
	(0.032)	(0.031)	(0.039)
log_size		0.432***	0.429***
		(0.033)	(0.033)
log_current_version		0.430***	0.428***
		(0.057)	(0.057)
log_last_updated		−0.909***	−0.879***
		(0.071)	(0.071)
is_free		1.418***	1.423***
		(0.050)	(0.051)
is_family_category		0.032	0.384
		(0.037)	(0.333)
is_game_category		0.624***	−2.374***
		(0.046)	(0.450)
is_tools_category		0.272***	−0.125
		(0.054)	(0.423)
rating:is_family_category			−0.085
			(0.080)
rating:is_game_category			0.702***
			(0.106)
rating:is_tools_category			0.095
			(0.104)
is_content_everyone		−0.067*	−0.066*
		(0.035)	(0.035)
Constant	3.822***	2.261***	2.469***
	(0.133)	(0.154)	(0.181)
Observations	5,408	5,408	5,408
R^2	0.028	0.265	0.271
Adjusted R^2	0.028	0.264	0.269
Residual Std. Error	1.143 (df = 5406)	0.995 (df = 5398)	0.991 (df = 5395)
F Statistic	155.421*** (df = 1; 5406)	216.073*** (df = 9; 5398)	166.929*** (df = 12; 5395)
Note:			*p<0.1; **p<0.05; ***p<0.01

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.

Because our data fails to meet the assumption of homoskedasticity, we adopt the large-sample assumptions (assumptions 1 and 2) instead. Specifically, we report robust standard errors rather than non-adjusted errors in our results.

2.3.0.2 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?

3 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?

4 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

5 Appendix