**STATISTICS AND ECONOMETRICS (MANM526)**

Word count: 3098

Student Name: Eldhose Varghese

Student URN: 6793260

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Sl\_No** | **Content** | **Page Number** |
| 1 | Introduction | 4 |
| 2 | Descriptive Analysis | 6 |
| 3 | Exploratory Analysis | 10 |
| 4 | Main Regression Analysis | 14 |
| 5 | Diagnostics and Robustness Analysis | 20 |

**List of Tables**

|  |  |  |
| --- | --- | --- |
| **Table Number** | **Table Name** | **Page Number** |
| 1 | Descriptive statistics | 6 |
| 2 | One-way ANOVA test | 8 |
| 3 | Z test | 8 |
| 4 | Correlation matrix | 9 |
| 5 | Baseline Model (Model 1) | 14 |
| 6 | Model 2 | 17 |
| 7 | Comparison of Model 1(baseline) and Model 3(Robust) | 21 |
| 8 | Model 4 | 22 |

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **Figure Number** | **Figure Name** | **Page Number** |
| 1 | distribution of app price(logged) | 10 |
| 2 | distribution of app rating | 11 |
| 3 | distribution of the app size | 11 |
| 4 | distribution of the number of languages | 12 |
| 5 | Boxplot of app revenue(logged) | 12 |
| 6 | Boxplot of app rating | 13 |
| 7 | Boxplot of the Number of Languages | 13 |
| 8 | Margins plot - Monetization Strategy | 16 |
| 9 | Margins plot -Target Age | 17 |
| 10 | Margins plot | 19 |
| 11 | Residual plot | 20 |
| 12 | Breusch-Pagan/Cook-Weisberg test | 20 |
| 13 | White's test | 20 |
| 14 | Scatter Plot | 23 |
| 15 | Quadratic Fit Plot | 23 |

**1. Introduction**

**1.1 Data**

The dataset comprises information on a sample of paid applications from the Google Play app store, specifically gathered during a single period in June 2023. These paid apps require users to pay a purchase price upfront before downloading and installing them on their devices. The dataset consists of 3,173 data points, each with 16 distinct attributes.

**1.2 Variables**

*Dependent variables*: This is the variable of primary interest. The dependent variable is the one we aim to predict or explain. It is depicted as the variable whose variation we are trying to understand based on other variables. We use app revenue (in USD) as the dependent variable to measure the financial performance of the apps. Revenue data frequently displays a skewed distribution, where a small number of apps generate considerably more income than the bulk. Logarithmic transformation aids in normalizing the skewness of the data, rendering it more symmetrical and appropriate for statistical analysis. Hence the app revenue is utilized in its natural logarithm-transformed form for all analyses.

*Independent variables*: These are the factors or predictor variables that can be used to predict or explain the outcome. In ordinary least squares (OLS) regression, the independent variables are employed to elucidate or forecast the dependent variable. They are regarded as the primary factors or determinants that we postulate to have an impact on the dependent variable. Independent variables might take the form of continuous, categorical, or a combination of both. The independent variables are the rating (average star rating given by users, a measure of app quality), app price (log-transformed to mitigate the skewness of the variable), monetization strategies (encoded), and target age (encoded).

*Control variables*: Control variables are a distinct category of independent variables. They are incorporated into the OLS regression model to mitigate the influence of factors that are not the main focus but could potentially distort the association between the independent and dependent variables. Through the process of controlling for these variables, we can isolate the precise influence of the independent variables on the dependent variable, guaranteeing a model that is more precise and dependable. The control variables are the number of languages for which the app is available and the app's main category(encoded).

**1.3 Objective**

The app store is a very competitive market. Implementing efficient tactics to differentiate oneself from the vast number of applications and attain substantial downloads and money is of paramount significance. This study aims to assess the efficacy of various competing strategies in achieving app success.

**1.4 Model**

The following model is used as the baseline model and is subjected to Ordinary Least Square (OLS) regression analysis.

Ln(revenue) = β0 + β1 × rating + β2 × Ln(price) + β3 × monetization\_strategies\_dum + β4 × age\_target \_dum + β5 × num\_langs + β6 × main\_category \_dum + ε

**2. Descriptive Analysis**

**2.1 Summary Statistics**

**Table 1** **- Descriptive statistics**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | (1) |  |  |  |  |
|  |  |  |  |  |  |
|  | count | mean | min | max | sd |
| 0 |  |  |  |  |  |
| app size | 1599 | 3.11e+07 | 1024 | 1.18e+09 | 5.81e+07 |
| app rating | 1599 | 4.072858 | 1 | 5 | .6682136 |
| Natural log of app revenue | 1599 | 12.9011 | 6.907755 | 19.70641 | 2.185424 |
| Natural log of app price | 1599 | 1.394806 | -2.995732 | 5.298267 | .8662884 |
| 1 |  |  |  |  |  |
| app size | 1185 | 8.38e+07 | 82944 | 1.61e+09 | 1.57e+08 |
| app rating | 1185 | 4.182616 | 1.7 | 5 | .4887086 |
| Natural log of app revenue | 1185 | 14.11787 | 7.600903 | 20.61987 | 2.197322 |
| Natural log of app price | 1185 | 1.240712 | -1.203973 | 3.496205 | .7563909 |
| Total |  |  |  |  |  |
| app size | 2784 | 5.35e+07 | 1024 | 1.61e+09 | 1.15e+08 |
| app rating | 2784 | 4.119576 | 1 | 5 | .6007849 |
| Natural log of app revenue | 2784 | 13.41902 | 6.907755 | 20.61987 | 2.271261 |
| Natural log of app price | 2784 | 1.329216 | -2.995732 | 5.298267 | .8246969 |
| Observations | 2784 |  |  |  |  |

**Non-Gaming Apps**

*Size*: The average size of these apps is approximately 31.1 million units. The size varies widely, as indicated by a large standard deviation.

*Rating*: The average rating is around 4.07, with all apps having at least a rating of 1, and some reaching the maximum of 5.

*Log\_Rev (Log of Revenue)*: These apps have an average logged revenue of about 12.9, indicating that there's a range in the revenues they generate.

*Log\_Price (Log of Price)*: The average logged price is 1.39, with some apps being free (as indicated by negative logged prices) and the highest prices reaching around 5.30.

**Gaming Apps**

*Size*: These apps are substantially larger on average, at about 83.8 million units, suggesting that game apps may require more storage space.

*Rating*: They have a slightly higher average rating than Non-gaming apps at approximately 4.18, which could suggest a slightly better user reception or satisfaction.

*Log\_Rev*: The average logged revenue is higher at about 14.1, suggesting that game apps potentially generate more revenue.

*Log\_Price*: The average logged price is slightly lower than Non-game apps at 1.24, which could indicate that game apps might be more affordable or more likely to be offered at lower prices.

**Total Sample**

*Size*: The combined average size is about 53.5 million units, which is between the average sizes of Non-game and Game apps.

*Rating*: The average rating for the total sample is 4.12, which is in line with the ratings for the individual categories, suggesting consistent rating scores across both game and non-game apps.

*Log\_Rev*: The average logged revenue for the entire sample is 13.42, closer to the average for game apps, potentially indicating a larger impact of game apps on total revenue.

*Log\_Price*: The overall average logged price is 1.32, showing that when combined, the pricing of apps does not significantly deviate from the averages of the individual categories.

**Interpretation and Implications**

*Size*: Game apps tend to be larger compared to non-game apps, which might reflect more content or higher-quality graphics commonly associated with games.

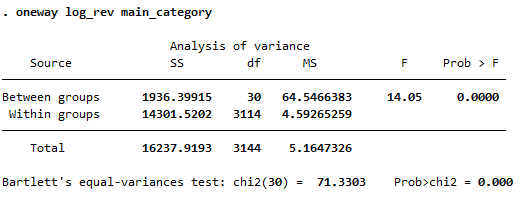
*Rating*: Both categories seem to perform well in terms of ratings, but game apps have a slightly higher average, indicating a favorable user perception.

*Revenue*: Game apps also appear to generate higher revenues, which could be due to in-app purchases or higher user engagement.

*Price*: The pricing strategy for game apps might be more competitive, given the slightly lower average log price compared to non-game apps.

**2.2 One-way ANOVA**

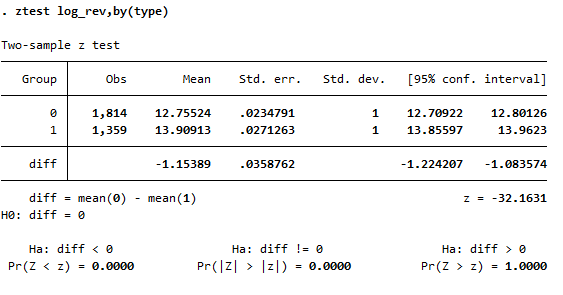
**Table 2 - One-way ANOVA test**



From the above test results, it is clear that the p-value is 0.0000, which is less than the significance level of 0.05. This means we can reject the null hypothesis that there is no difference in average logged revenue across the different categories. In conclusion, there is a statistically significant difference in app revenue (logged) across the main categories.

**2.3 Z Test**

**Table 3 – z test**



From the above table, it is clear that the p-value for the z-test (Pr(|Z| > |z|)) is 0.0000, which is less than the significance level of 0.05, indicating that the difference in means is statistically significant. In summary, there is a statistically significant difference in app revenue(logged) between games and non-games, with games having higher logged revenue on average. The test is very significant with a p-value of 0.0000, which is much lower than the conventional alpha level of 0.05. This means that we would reject the null hypothesis of no difference in average logged revenue between games and non-games apps.

**2.4 Correlation**

**Table 4 - Correlation matrix**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | (1) |  |  |  |  |
|  |  |  |  |  |  |
|  | Natural log of app revenue | app size | app rating | number of languages | Natural log of app price |
| Natural log of app revenue | 1 |  |  |  |  |
| app size | 0.0871\*\*\* | 1 |  |  |  |
| app rating | 0.202\*\*\* | 0.00675 | 1 |  |  |
| number of languages | 0.237\*\*\* | 0.0129 | 0.110\*\*\* | 1 |  |
| Natural log of app price | 0.0356 | 0.127\*\*\* | 0.0407\* | 0.0167 | 1 |
| Observations | 2757 |  |  |  |  |

\* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001

*log\_rev and size (0.0871)*: A positive but very weak correlation, suggesting that larger app sizes are marginally associated with higher logged revenues.

*log\_rev and rating (0.202)*: A weak positive correlation, indicating that higher ratings are slightly associated with higher logged revenues.

*log\_rev and num\_langs (0.237)*: This is weak to moderate positive correlation, which could mean that apps supporting more languages may see slightly higher logged revenues.

*log\_rev and log\_price (0.0356)*: This very weak positive correlation suggests that there is almost no linear relationship between the price of an app and its revenue when both are logged.

*size and rating (0.00675)*: This negligible correlation implies that there is no meaningful association between an app’s size and its rating.

*size and num\_langs (0.0129)*: Similarly, this shows a very weak positive correlation, suggesting that there's barely any relationship between the size of an app and the number of languages it supports.

*size and log\_price (0.127)*: A weak positive correlation, which might indicate that larger apps tend to be a bit more expensive when the price is logged.

*rating and num\_langs (0.110)*: A weak positive correlation suggests that apps with higher ratings might support a greater number of languages, although the relationship is not strong.

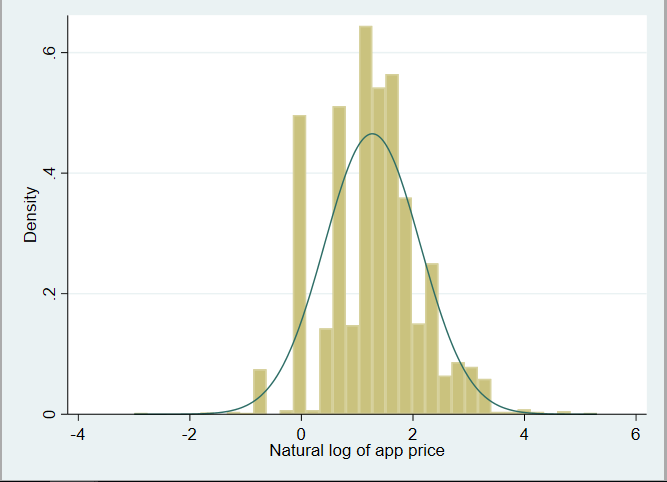
*rating and log\_price (0.0407)*: Another very weak positive correlation, indicating almost no linear relationship between an app's rating and its logged price.

*num\_langs and log\_price (0.0167)*: This very weak positive correlation suggests no real linear relationship between the number of languages an app supports and its logged price.

**3. Exploratory Analysis**

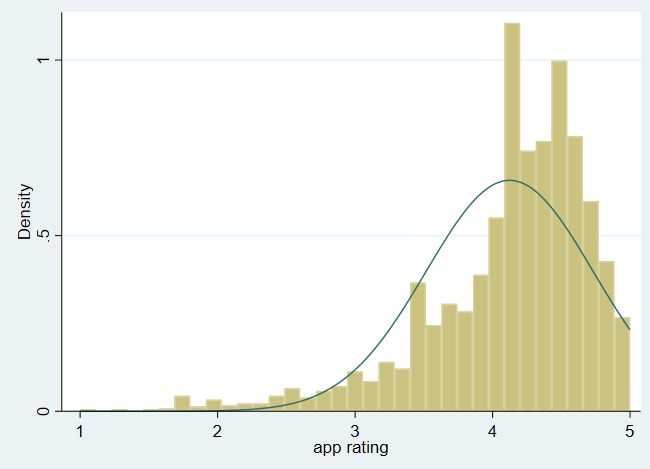
**3.1 Distribution or Skewness of Variables**

**Figure 1 – distribution of app price(logged)**

****

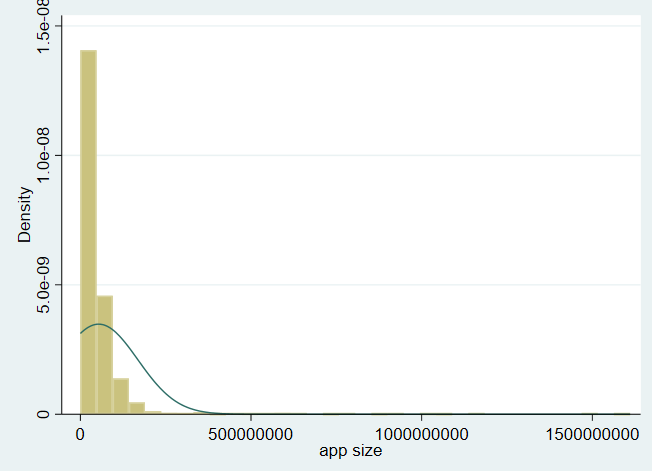
The distribution of the natural log of app prices is somewhat skewed to the right, indicating that there are more apps with lower prices and fewer with higher prices.

**Figure 2 –** **distribution of app rating**

****

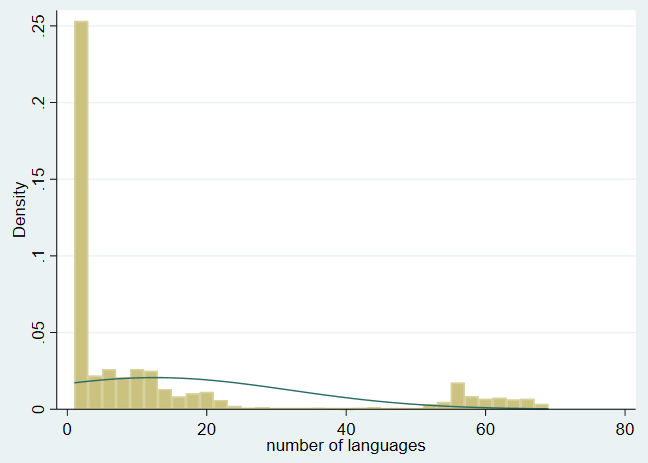
The distribution of app ratings is left-skewed, meaning there are fewer low-rated apps and a higher concentration of high-rated apps. There is a significant concentration of apps with ratings between 4 and 5, as indicated by the height of the histogram bars and the peak of the density curve in that range.

**Figure 3 – distribution of the app size**

****

The distribution of app size is extremely right-skewed, with the vast majority of apps being of smaller size. There are very few apps in the dataset that have a large size which could represent a separate category of app types (e.g., games with extensive graphics or apps with built-in media content).

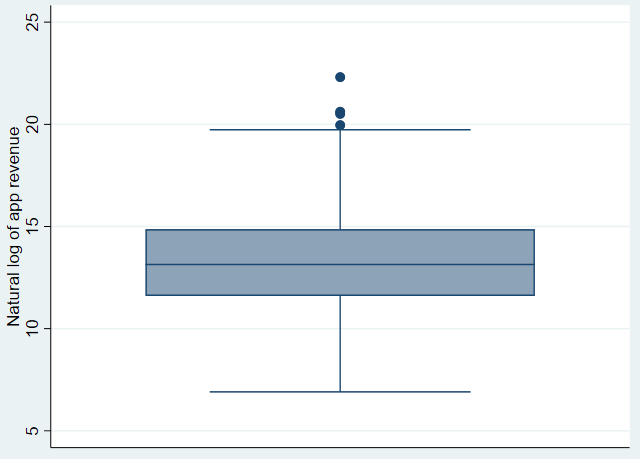
**Figure 4 – distribution of the number of languages**

****

The distribution is heavily right-skewed, indicating that most apps support only a few languages. The long tail to the right indicates that there are relatively few apps that support a large number of languages

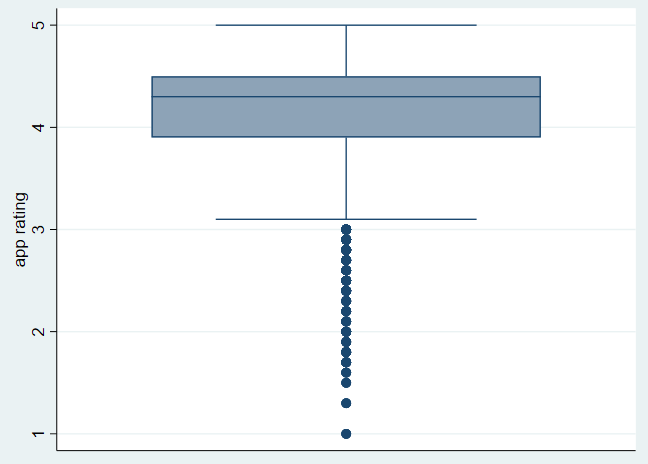
**3.2 Outliers**

**Figure 5 – Boxplot of app revenue(logged)**

****

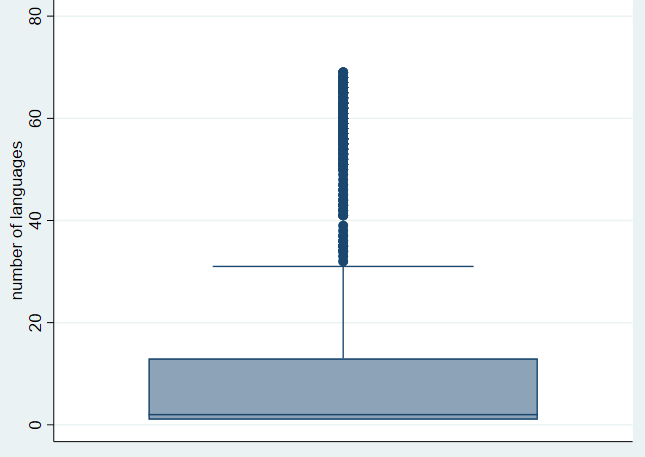
The presence of outliers above the upper whisker suggests that there are a few apps with exceptionally high revenue when compared to the rest of the data.

**Figure 6 – Boxplot of app rating**



The median rating is above 4, indicating that over half of the apps have ratings above this value. There is a considerable number of outliers below the lower whisker, indicating that there are several apps with ratings substantially lower than the median. These could be considered unusually low-rated apps compared to the rest. In summary, the box plot indicates that, while most apps have high user ratings, there's a set of apps with significantly lower user satisfaction.

**Figure 7 – Boxplot of the Number of Languages**

****

The numerous outliers indicate that while most apps support a lower number of languages, several apps support a significantly higher number of languages which may suggest that there is a subset of apps designed to reach a wide, multilingual audience.

**4. Main Regression Analysis**

**Table 5 - Baseline Model (Model 1)**

|  |  |
| --- | --- |
|  | (1) |
|  | Natural log of app revenue |
| app rating | 0.521\*\*\* |
|  | (0.061) |
|  |  |
| Natural log of app price | 0.148\*\*\* |
|  | (0.044) |
|  |  |
| paid | 0.000 |
|  | (.) |
|  |  |
| paid;ads | 0.456\* |
|  | (0.202) |
|  |  |
| paid;inapp | 1.172\*\*\* |
|  | (0.100) |
|  |  |
| Everyone | 0.000 |
|  | (.) |
|  |  |
| Everyone 10+ | 0.795\*\*\* |
|  | (0.148) |
|  |  |
| Mature 17+ | 0.788\*\*\* |
|  | (0.231) |
|  |  |
| Teen | 0.782\*\*\* |
|  | (0.119) |
|  |  |
| number of languages | 0.024\*\*\* |
|  | (0.002) |
|  |  |
| Main\_category dummies | Included |
|  |  |
|  |  |
|  |  |
| Constant | 9.763\*\*\* |
|  | (0.430) |
| Observations | 3142 |
| *R*2 | 0.250 |
| Adjusted *R*2 | 0.241 |

Standard errors in parentheses

\* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001

**4.1 Model Fit**

*R-squared*: The model has an *R*2 of 0.2503, which means that approximately 25.03% of the variability in the logged app revenue can be explained by the independent variables included in the model.

*F-statistic*: The F-statistic is 27.26, and it is highly significant (Prob > F = 0.0000), suggesting that the model as a whole is statistically significant and that the independent variables, taken together, do have a relationship with the logged app revenue.

**4.2 Statistical Significance of Coefficients, and the Effect Size**

*App Rating*: The coefficient for app rating is 0.521, and it's highly statistically significant (p < 0.001). This implies that holding all other variables constant, a one-unit increase in app rating is associated with an increase of 52.1% in revenue on average.

*App Price*: The coefficient for the app price(logged) is 0.148 and is statistically significant (p = 0.001). This means that a one percent increase in app price leads to a 0.148% increase in revenue.

*“Paid; Ads” Monetization Strategy*: The “paid;ads” monetization strategy has a coefficient of 0.456, indicating a significant increase in revenue associated with this strategy compared to the “paid” monetization strategy. Applying the “paid;ads” monetization strategy yields 45.6% more revenue compared to using the “paid” monetization strategy on average at 5% significance.

*“Paid; Inapp” Monetization Strategy*: On average, employing a monetization strategy that combines paid access with in-app purchases is associated with a 117.2% increase in revenue compared to a strategy that solely relies on upfront payment, with this result being statistically significant at the 5% level.

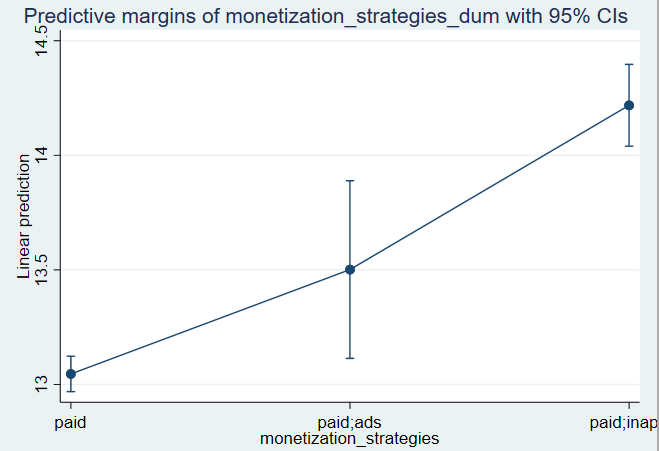
*Age Target*: Apps targeted at "Everyone 10+", "Mature 17+", and "Teen" have positive and significant coefficients, suggesting these age targets are associated with higher revenues compared to the “Everyone” age category. Apps targeting the "Everyone 10+" age category yield 79.5% more revenue compared to apps that target the “Everyone” age category on average at 5% significance.

*Number of Languages*: The coefficient for the number of languages is 0.024, which is statistically significant (p < 0.001), indicating that supporting additional languages is positively associated with app revenue. Each additional language provided leads to a 2.4% increase in app revenue on average at 5% significance.

*Main Category*: Most category dummy variables are not statistically significant, suggesting that they don't have a distinct effect on the logged revenue when controlling for other factors.

**4.3 Monetization Strategy**

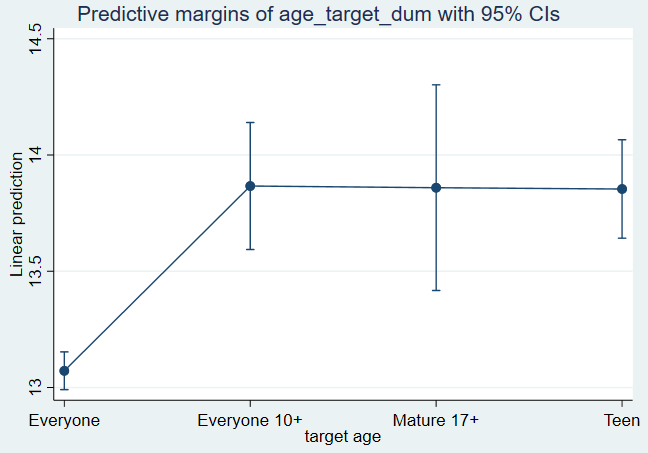
**Figure 8 – Margins plot - Monetization Strategy**



From the above graph, we can infer that the "paid; in-app" monetization strategy is associated with the highest revenue generation among the three strategies analyzed.

**4.4 Target Age Level**

**Figure 9 –** **Margins plot - Target Age**



From the margins plot above it is clear that targeting specific age levels such as "Everyone 10+," "Mature 17+," or "Teen" appears to yield a higher revenue compared to targeting everyone.

**4.5 Differential effect of app rating on app revenue for different monetization strategies**

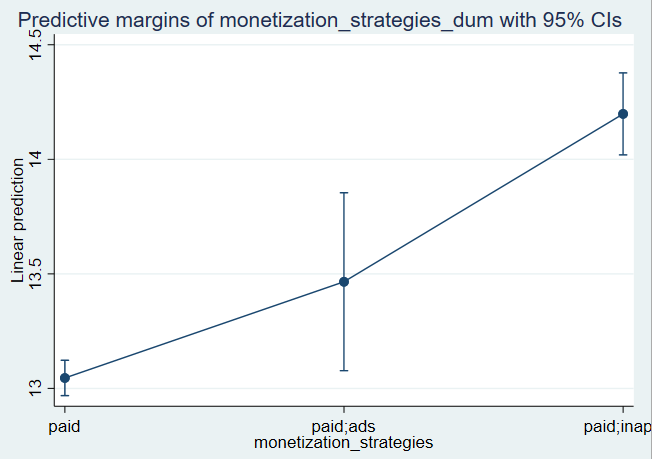
**Table 6 - Model 2**

|  |  |
| --- | --- |
|  | (1) |
|  | Natural log of app revenue |
| app rating | 0.489\*\*\* |
|  | (0.065) |
|  |  |
| Natural log of app price | 0.149\*\*\* |
|  | (0.044) |
|  |  |
| paid | 0.000 |
|  | (.) |
|  |  |
| paid;ads | 3.879\* |
|  | (1.610) |
|  |  |
| paid;inapp | -0.744 |
|  | (0.785) |
|  |  |
| Everyone | 0.000 |
|  | (.) |
|  |  |
| Everyone 10+ | 0.805\*\*\* |
|  | (0.148) |
|  |  |
| Mature 17+ | 0.767\*\*\* |
|  | (0.231) |
|  |  |
| Teen | 0.789\*\*\* |
|  | (0.119) |
|  |  |
| paid # app rating | 0.000 |
|  | (.) |
|  |  |
| paid;ads # app rating | -0.839\* |
|  | (0.391) |
|  |  |
| paid;inapp # app rating | 0.460\* |
|  | (0.187) |
|  |  |
| number of languages | 0.024\*\*\* |
|  | (0.002) |
|  |  |
| Main\_category dummies | Included |
|  |  |
|  |  |
|  |  |
| Constant | 9.913\*\*\* |
|  | (0.442) |
| Observations | 3142 |
| *R*2 | 0.253 |
| Adjusted *R*2 | 0.243 |

*Paid;ads # app rating*: Coefficient = -0.839, p < 0.05. The negative coefficient indicates that the beneficial effect of app rating on revenue is less for apps that use the paid-with-ads strategy. For one unit increase in app rating, there is an 83.9% decrease in revenue for “Paid;ads” monetization strategy compared to the “paid” strategy on average at 5% significance.

*Paid;inapp # app rating*: Coefficient = 0.460, p < 0.05. This positive coefficient indicates that the beneficial effect of app rating on revenue is more for apps that use the paid with in-app purchases strategy. This means that as the quality (rating) of the app increases, the revenue differential for this strategy increases as well. For one unit increase in app rating, there is a 46% increase in revenue for “Paid;inapp” monetization strategy compared to the “paid” strategy on average at 5% significance.

**Figure 10 – Margins plot**

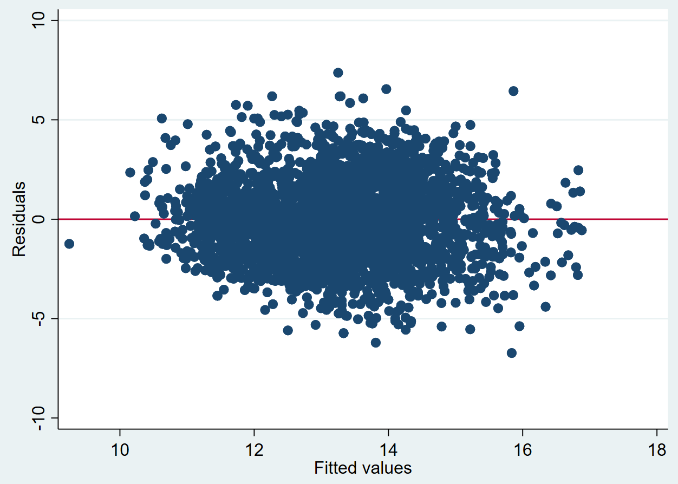
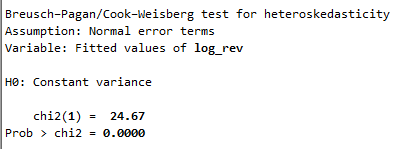
****

The best monetization strategy for high-quality apps (those with higher ratings) appears to be paid with in-app purchases (paid;inapp), as it has the highest point estimate for the predicted revenue.

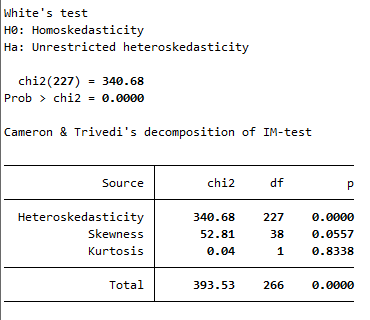
The worst strategy appears to be the baseline Paid-only (paid) strategy, as it has the lowest point estimate for the predicted revenue.

**5. Diagnostics and Robustness Analysis**

**Figure 11 - Residual plot Figure 12 - Breusch-Pagan/Cook-Weisberg test**

** **

**Figure 13 - White's test**



From Fig 11 the spread of the residuals appears to increase as the fitted values increase. This indicates heteroscedasticity, meaning that the variance of the errors is not constant across all levels of the independent variable. This plot shows some signs of a non-random pattern, specifically the funnel shape where the residuals spread out with larger fitted values.

From Fig 12 it is clear that the p-value is less than any conventional significance level (0.05, 0.01, etc.), we reject the null hypothesis of constant variance in favor of the alternative hypothesis that the variance of the residuals is not constant (heteroskedasticity is present).

From Fig 13 it is clear that the p-value is zero, so we reject the null hypothesis in favor of the alternative hypothesis of unrestricted heteroskedasticity. Therefore the test confirms that heteroskedasticity is present. An appropriate remedy would be to use robust standard errors to correct for the heteroskedasticity.

**Table 7 – Comparison of Model 1(baseline) and Model 3(Robust)**

|  |  |  |
| --- | --- | --- |
|  | (Model 1) | (Model 3) |
|  | Natural log of app revenue | Natural log of app revenue |
| app rating | 0.521\*\*\* | 0.521\*\*\* |
|  | (0.061) | (0.067) |
|  |  |  |
| Natural log of app price | 0.148\*\*\* | 0.148\*\* |
|  | (0.044) | (0.046) |
|  |  |  |
| paid | 0.000 | 0.000 |
|  | (.) | (.) |
|  |  |  |
| paid;ads | 0.456\* | 0.456\* |
|  | (0.202) | (0.221) |
|  |  |  |
| paid;inapp | 1.172\*\*\* | 1.172\*\*\* |
|  | (0.100) | (0.098) |
|  |  |  |
| Everyone | 0.000 | 0.000 |
|  | (.) | (.) |
|  |  |  |
| Everyone 10+ | 0.795\*\*\* | 0.795\*\*\* |
|  | (0.148) | (0.153) |
|  |  |  |
| Mature 17+ | 0.788\*\*\* | 0.788\*\* |
|  | (0.231) | (0.299) |
|  |  |  |
| Teen | 0.782\*\*\* | 0.782\*\*\* |
|  | (0.119) | (0.127) |
|  |  |  |
| number of languages | 0.024\*\*\* | 0.024\*\*\* |
|  | (0.002) | (0.002) |
|  |  |  |
| Main\_category dummies | Included | Included |
|  |  |  |
|  |  |  |
|  |  |  |
| Constant | 9.763\*\*\* | 9.763\*\*\* |
|  | (0.430) | (0.377) |
| Observations | 3142 | 3142 |
| *R*2 | 0.250 | 0.250 |
| Adjusted *R*2 | 0.241 | 0.241 |

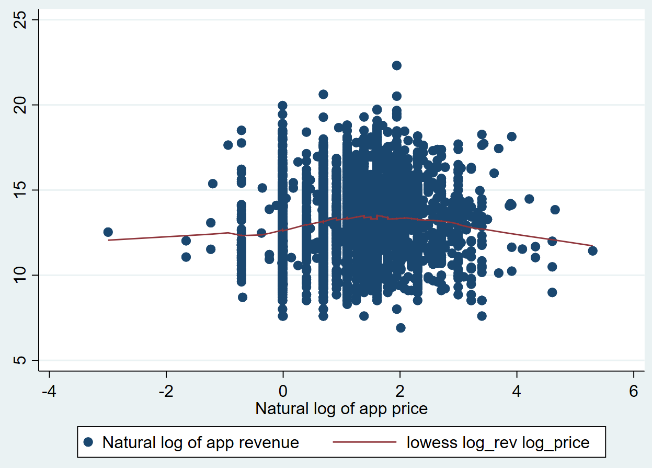
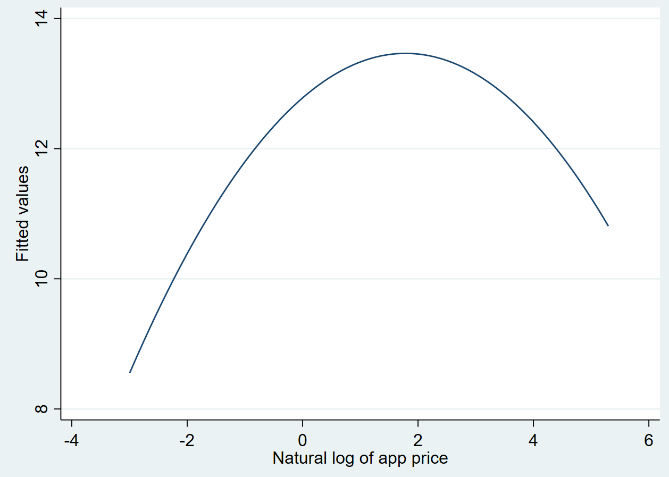
Both models seem to have identical coefficients for each predictor while the standard errors for model 3 are improved and are robust to heteroskedasticity.

**5.2 Quadratic effect of the app price (logged) on the app revenue**

**Table 8 – Model 4**

|  |  |
| --- | --- |
|  | (1) |
|  | Natural log of app revenue |
| app rating | 0.517\*\*\* |
|  | (0.061) |
|  |  |
| Natural log of app price | 0.511\*\*\* |
|  | (0.093) |
|  |  |
| log\_price\_sq | -0.136\*\*\* |
|  | (0.031) |
|  |  |
| paid | 0.000 |
|  | (.) |
|  |  |
| paid;ads | 0.454\* |
|  | (0.201) |
|  |  |
| paid;inapp | 1.185\*\*\* |
|  | (0.099) |
|  |  |
| Everyone | 0.000 |
|  | (.) |
|  |  |
| Everyone 10+ | 0.822\*\*\* |
|  | (0.148) |
|  |  |
| Mature 17+ | 0.822\*\*\* |
|  | (0.230) |
|  |  |
| Teen | 0.822\*\*\* |
|  | (0.119) |
|  |  |
| number of languages | 0.024\*\*\* |
|  | (0.002) |
|  |  |
| Main\_category dummies | Included |
|  |  |
|  |  |
|  |  |
| Constant | 9.619\*\*\* |
|  | (0.430) |
| Observations | 3142 |
| *R*2 | 0.255 |
| Adjusted *R*2 | 0.246 |

**Figure 14 - Scatter Plot** **Figure 15 - Quadratic Fit Plot**

From Figure 14 it is clear that the curve appears relatively flat at the lower end of app prices and then starts to rise as the price increases, peaking at around the natural log of price equal to 2. After this peak, the curve begins to decline, indicating that beyond a certain price point, increasing the price is associated with a decrease in revenue. Figure 15 depicts a quadratic relationship in the curve. The value rises, attains its highest point (the apex of the parabola), and thereafter declines. This demonstrates a quadratic relationship, in which there exists optimal pricing that maximizes revenue. The results from both curves indicate that there is an optimal price point for apps that maximizes revenue. Setting the price too low might increase the number of installs but could lead to lower overall revenue. Conversely, setting the price too high might boost revenue per install but could lead to a decrease in the total number of installs, also reducing overall revenue.

The peak of the quadratic curve represents the optimal price point. Below this price, revenue can be increased by raising the price, and above this price, revenue can be increased by lowering the price. For revenue generation, it is better to set the price somewhere in the middle range - the point at which the quadratic curve reaches its maximum. This optimal pricing strategy balances the trade-off between the number of installs and the revenue per install to achieve maximum overall revenue.

**5.3 Potential Endogeneity Problems in the Baseline Model**

*Omitted Variable Bias*: If important variables that affect app revenue are omitted from the model, the estimated effects of the included variables may be biased. For example, the size of an app could influence both the app rating and the app revenue—larger apps might offer more features, leading to higher ratings and revenue. If app size is correlated with the included variables and influences revenue, its omission could bias the estimated coefficients. Similarly, the age of an app could affect its revenue, as older apps might have more reviews and a more established user base. Including the release date could control for the lifecycle effects of an app.

*Measurement Error*: If there is an error in measuring the included variables, such as app rating or price, this could result in biased and inconsistent estimates.

*Simultaneity:* The pricing of an application may be determined based on anticipated ratings, and conversely, the ratings may be influenced by the app's price. If developers modify prices in reaction to ratings or the anticipated outcomes of those ratings, this simultaneous relationship could result in endogeneity.

Using additional variables in the dataset could improve the model by reducing omitted variable bias:

App Size: Larger apps may offer more content or functionality, which could be associated with higher revenue. Including app size could help control for the scope and quality of the app.

App Release Date: Including the release date could control for app maturity. Older apps may have had more time to accumulate a user base and revenue.

Developer Name: If certain developers have a track record of creating successful apps, their names could be associated with higher app revenue. Including a variable for the developer might control for brand effects.

Collecting panel data could significantly enhance the model by:

*Controlling for Unobserved Heterogeneity*: With panel data, fixed effects models can be used to control for time-invariant unobserved factors that might affect revenue, such as developer skill or app quality.

*Addressing Simultaneity*: Panel data allows for the use of instrumental variables (IV) or differences-in-differences (DiD) approaches, which can help address simultaneity by exploiting exogenous variation over time.

*Improving Causal Inference*: By observing changes over time, panel data can help establish temporal precedence and potentially support causal interpretations of the relationships between variables.