**STATISTICS AND ECONOMETRICS MANM526**

**AMRUTHA JEEVA SATHEESAN**

**6807428**

**Total wordcount: 3362**

**1. Introduction:**

The given dataset is a panel dataset from the period of June 2023 with variables related to a sample of paid applications provided by Google Play app store. Users are needed to pay the app price before they can download and install the app from play store. There are a total of 16 variables and 3,173 observations in the given dataset.

The primary purpose of this project is to conduct a descriptive and exploratory analysis and Ordinary Least Square Regression to calculate the effect of app rating, app price (logged), monetization strategies, and age target on app revenue (logged) using control variables as the number of available languages and app main category. The variables price and revenue have been logged which helps to normalise the data. Revenue (logged) is the dependant variable which represents the target variable that we are going to predict. Dependant variable is also called explained or response variable. The independent variables are the variables that have an effect on the dependant variable. They are also called the regressors or explanatory variables. In our model, the regressors are rating, price (logged), monetization strategies, and age target. Rating is a measure of the quality of the application. The more rating, the higher the app quality. Price is the price of application in USD. Monetization strategies identifies the strategy for revenue generation. Here, there are two strategies-: in-app purchases and/or in-app advertisement along with app price. Age target is used to identify the age category target of the app. The control variables num\_langs and main\_category provides the number of languages the app is available and app category like games, music, books etc respectively. They are used to improve the accuracy and reliability of the model. Apart from these, there are other general variables like product id, name, developer, version, release date etc.

Model:

The baseline model is the foundation for the analysis which includes the independent variables, dependent variable, and control variables. In our scenario, the baseline model may be represented as follows:

Revenue (logged)=β0 + β1×Rating + β2× Price (logged) + β3×Monetization Strategies + β4×Age Target + β5×Number of Available Languages + β6×App Main Category + *ϵ*

β0 : Intercept term

β1, β2, β3, β4, β5, β6 : Co-efficients of independent and control variables which represents their impact on the dependant varible

*ϵ* : error term which contains all other factors affecting Revenue(logged)

**2. Descriptive analysis:**

A new variable sub\_category has been generated to categorize main category of apps to games (1) or non-games (0). New dummy variables are generated for categorical variables which will be included in the regression baseline model—monetization strategy (monet\_strat\_dum), age target (age\_target\_dum), and main category (main\_cat\_dum), to intensify the analysis of their impact on the dependent variable.

2.1 Two-way Table:

The two-way table below (Table 1) provides a summary statistic of the numerical variables for category of games (1), category of non-games (0) and full sample.

It is observed that number of non-games apps exceeds the number of gaming apps by almost 400. Both game and non-game categories have almost same mean price(logged) with approximately 1.2 indicating a moderate price range. Mean revenue (logged) is slightly higher for game apps with 14.11 compared to non-game apps with 12.88 which explains more revenue is generated from game apps, with the maximum value going up to 20.61 . There is a good number of languages in which two categories of apps are available, with average around 11 languages for game apps and 15 for non-game apps, which indicates the level of internationalisation and broad accessibility of the apps. The maximum number of languages for each category goes up to 69. Both categories have a high average rating of above 4 given by it’s users which is an excellent indication of user satisfaction and usability of apps. A difference in the mean size of game and non-game apps can be noted, with game apps having an average size of 8.38 MB while non-game apps have only 3.08 MB. The reasons of game apps having more size can be due to the additional graphics, complexity and special features like multiple game levels, animations, sound effects etc which non-game apps may not have.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | (1) |  |  |  |  |
|  |  |  |  |  |  |
|  | count | mean | min | max | sd |
| 0 |  |  |  |  |  |
| Log of Price | 1572 | 1.398333 | -2.995732 | 5.298267 | .8673919 |
| Log of revenue | 1572 | 12.88225 | 6.907755 | 19.70641 | 2.175408 |
| No. of languages | 1572 | 15.05662 | 1 | 69 | 22.41246 |
| App Rating | 1572 | 4.074682 | 1 | 5 | .6665474 |
| App Size | 1572 | 3.08e+07 | 1024 | 1.18e+09 | 5.82e+07 |
|  |  |  |  |  |  |
| 1  Log of Price | 1185 | 1.240712 | -1.203973 | 3.496205 | .7563909 |
| Log of revenue | 1185 | 14.11787 | 7.600903 | 20.61987 | 2.197322 |
| No. of languages | 1185 | 11.08439 | 1 | 69 | 15.37421 |
| App Rating | 1185 | 4.182616 | 1.7 | 5 | .4887086 |
| App Size | 1185 | 8.38e+07 | 82944 | 1.61e+09 | 1.57e+08 |
| Total |  |  |  |  |  |
| Log of Price | 2757 | 1.330585 | -2.995732 | 5.298267 | .8250778 |
| Log of revenue | 2757 | 13.41334 | 6.907755 | 20.61987 | 2.268514 |
| No. of languages | 2757 | 13.34929 | 1 | 69 | 19.79265 |
| App Rating | 2757 | 4.121074 | 1 | 5 | .5989304 |
| App Size | 2757 | 5.36e+07 | 1024 | 1.61e+09 | 1.15e+08 |
| Observations | 2757 |  |  |  |  |

**Table 1: Two-way table**

2.2 Z-TEST :

An z-test is used to compare revenue(logged) between sub\_categories 0 (non-game apps) and 1 (game apps). It can be inferred that there is a significant difference in mean revenue(logged) between the two sub\_categories. Group 1, representing game apps, has a significantly higher mean revenue(logged) (13.90913) compared to sub-category 0 (12.75524). The null hypothesis states that there is no difference between the means of the two sub\_categories and the alternative hypothesis states that there is significant difference among the means of two sub\_categories. The z-test statistic of -32.1631 helps to understand that the null hypothesis can be rejected. That means, game apps demonstrate a statistically higher mean revenue (logged) than non-game apps, at a significance level of 5% , as indicated by the negative difference of -1.15389.

A screenshot of a computer

Description automatically generated

**Table 2: Z-Test to compare means of game and non-game groups**

2.3 ONE-WAY ANOVA:

The one-way ANOVA table explain significant differences among the groups (F(30, 3114) = 14.05, p = 0.0000). The high F-statistic depicts the variation in means of groups. However, Bartlett's test indicates dissimilar variances (chi2(30) = 71.3303, p = 0.000), highlighting the need for cautious analysis. The between-group sum of squares is 1936.39915, within-group is 14301.5202, and the total sum of squares is 16237.9193. These findings marks the importance of considering both overall group differences and potential variations in variances for accurate interpretation.

A screenshot of a graph

Description automatically generated

**Table 3: One-way ANOVA table to find statistical significance among groups**

2.4 Correlation Matrix:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | (1) |  |  |  |  |
|  |  |  |  |  |  |
|  | Log of revenue | App Size | App Rating | No. of languages | Log of Price |
| Log of revenue | 1 |  |  |  |  |
| App Size | 0.0871\*\*\* | 1 |  |  |  |
| App Rating | 0.202\*\*\* | 0.00675 | 1 |  |  |
| No. of languages | 0.237\*\*\* | 0.0129 | 0.110\*\*\* | 1 |  |
| Log of Price | 0.0356 | 0.127\*\*\* | 0.0407\* | 0.0167 | 1 |
| Observations | 2757 |  |  |  |  |

A correlation matrix is used to find the relation between the regressors/control variables and the dependant variable. Only numerical variables price (logged), number of languages, size and rating are used to find the strength of relation with dependant variable revenue(logged). As Table 3 depicts, correlation co-efficients of all the independent and control variables at different significant levels. Notably, there is a positive moderate correlation between

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

**Table 4: Correlation Matrix**

revenue(logged) and rating (0.202) at 0.001 significant level and with number of languages (0.237). Higher revenue generation is from these two variables as there is more tendency for the users to check the app rating, which is a measure of customer satisfaction, before installing the app. When the number of languages increases, there is more chance of people from different regions using the app. There is weak relation with size (0.0871) and price(logged) (0.0356), explaining the fact that size and app price have no much influence on the revenue generated .

**3. Exploratory Analysis:**

In this section, different graphs and visuals are used to analyse and interpret data, pre-check outliers, check the skewness or distribution of data and to visualise the relation between independent and dependant variables.

The below bar graph (fig 1.) illustrates mean logged revenue across three monetization strategies: "Paid In-App" which generates the highest mean revenue (logged) with 14.4901, followed by "Paid Ads" at 13.5402, and "Paid" at 12.9918, showcasing the varying revenue performance among the strategies. Some of the reasons users prefer paid in-app strategy are due to its free initial access to the app, flexibility in spending money according to their preferences, customization options and ad-free experience compared to ‘paid ads’ strategy.

A graph of a graph with numbers and text

Description automatically generated with medium confidence

**Fig 1. : Log\_revenue across different monetization strategies**

The four different age group categories are -Everyone, Everyone 10+, Mature 17+ and Teen. The pie chart (fig 2.) presented below clearly depicts the percentage of revenue from app for each age category.

The majority, accounting for 76.47% of revenue, is generated from apps designed for users of all age groups.13.08% revenue is from the apps targeting teens. Additionally, apps tailored for users over 10 years old, contribute 7.702% to the total revenue share. The smallest segment comes from apps made for mature users aged over 17, contributing 2.751% to the total revenue. The main reasons for apps for mature people contributing to the revenue generation can be due to their time constraints, more focus on work and family and privacy and security concerns.

A pie chart with different colored circles

Description automatically generated

**Fig 2. Log\_revenue over various age groups**

The box plot (fig 3.) analysis below of revenue (logged) across different devices reveals that the median revenue is highest for TV: Handheld devices, followed by Handheld:TV, and the lowest for Handheld devices. Also. There are some outliers for the revenue from handheld devices. Reasons for more revenue generation from TV:Handheld devices can be due to enhanced viewing experiences, different connectivity options, versatility etc.

A graph of a diagram

Description automatically generated with medium confidence

**Fig 3. Box plot of different device types**

When the revenue generation for devices across each sub category-Games and Non Games are analysed (fig 4.), it is significant that there is a good revenue from TV:Handheld devices with gaming apps, with a mean revenue(logged) value of 16.3573 compared to others, which have an approximate of around 12.

A graph with green bars

Description automatically generated

**Fig 4. Devices and revenue across non-game apps(0) & game apps(1)**

The distribution of revenue (logged) as depicted in the below histogram (fig 5.) illustrates a normal distribution characterized by a mean value of 13.24946 and standard deviation of 2.27543. This shows a balanced and symmetrical spread of revenue values.

The distribution of app rating (fig 6.) has skew towards left, indicating more concentration of rating towards the higher end (right side). The mean rating is 4.1237, suggesting that a good number of apps tend to receive higher rating, contributing to the leftward skew of the distribution.

A graph with a line

Description automatically generated

**Fig 5. Log\_revenue distribution**

A graph with a line going up

Description automatically generated

**Fig 6. Distribution of app rating**

In the scatter plot (fig 7.) of revenue (logged) vs price (logged), more concentration of revenue generation is observed for apps with price (logged) between 0 to 3. As the fitted line goes from left to right, it has a slight increase in the slope positively, indicating the positive relation between price(logged) and revenue(logged). The mean of price(logged) 1.27172 suggests that, on average, price of apps are moderate. The mean revenue (logged) of 13.24946 indicates a comparatively high revenue, implying that apps within this price range are successful in generating substantial revenue.

A graph of a graph showing the amount of revenue

Description automatically generated with medium confidence

**Fig 8. Relation between log\_price & log\_revenue**

**4. Main Regression Analysis:**

Ordinary Least Squares (OLS) regression is a statistical technique used modelling the relationship between dependent variable and one or more independent variables.

4.1 Baseline model:

In our model, the baseline model for OLS regression is;

Revenue (logged)=β0 + β1×Rating + β2× Price (logged) + β3×Monetization Strategies + β4×Age Target + β5×Number of Available Languages + β6×App Main Category + *ϵ*

Model-Fit Table:

In the below table 5., the Model-Fit table gives key metrics that evaluates how well the model fits the data. The results depicts a significant model (F(38, 3103) = 27.26, p < 0.0001), describing 25.03% of the variance in log\_revenue. The model uses 38 predictors which are significant in predicting the outcome from a statistical perspective. The R-squared value of 0.2503 recommends a moderate proportion of the variability is being explained. The adjusted R-squared value of 0.2411 showing that approximately 24.11% of the variability in the dependent variable is explained by the included independent variables, adjusting for the number of predictors. The Root Mean Squared Error (RMSE) is 1.9804., which is a measure of average magnitude of errors between actual and predicted values. The lower the PMSE value, the better the model is.

ANOVA Table:

The significance of our OLS regression model is explained by the ANOVA table. The model's sum of squares (SS) is 4062.62, which indicates the variability explained by the model, with 38 degrees of freedom (df), resulting in a mean square (MS) of 106.91, which asses the contribution of each predictor. The residual SS is 12170.13 which represents the unexplained variability in the model, df of 3103, and an MS of 3.92. The Total SS is 16232.75 with 3141 df and an MS of 5.17.

Interpretation of effect size:

Notable findings from the table 5. include analysing how the independent variables change with dependent variables, which can be interpreted from the co-efficient value. For one unit increase in rating, there is an increase of 52.1% in revenue. One percent increase in app price results in 0.14% increase in revenue. 2.4% increase in revenue will be there for one unit increase in number of languages.

All variables where p-value < 0.05, which is the significance level, are considered statistically significant. After thorough analysis, it can be found that most of the apps under main category are statistically not significant. Under main category, Art and Design is considered as the base category. Co-efficient value of 0.9173225 for Game apps suggests that Game apps tend to generate a revenue which is 0.91 units higher compared to Art and Design apps.

|  |  |
| --- | --- |
|  | (1) |
|  | Log of revenue |
| App Rating | 0.521\*\*\* |
|  | (0.061) |
|  |  |
| Log of 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) |
|  |  |
| No. 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

**Table 5: Regression table of baseline model**

4.1.1

By analysing the regression results of different monetization strategies, paid:inapp strategy generates the highest revenue, which is 1.172415 times compared to base category – paid. From the marginsplot (fig 9.) also, it can be confirmed that paid;inapp strategy has the highest predicted log\_revenue of 14.21826, followed by paid;ads with a predicted value of 13.5014. the lowest predicted revenue is from paid apps, with a value of 13.04585. the reasons behind highest revenue generation by paid;inapp could be due to the fact that it gives more opportunities for the users to spend money in the app only if they want to. Paid;ads apps has a revenue generation which is 0.455554 time the paid apps, which could be because the users need not pay any extra money for the ads, but while watching those ads, more revenue is generated for the app company.

A graph with a line

Description automatically generated

**Fig 9: Marginsplot for Monetization Strategy and Log\_revenue**

4.1.2

The regression results for age target groups depicts the coefficients and significance levels for "Everyone 10+," "Mature 17+," and "Teen" groups compared to the base category "Everyone." All three age groups have a positive coefficient, which states that they are related with higher predicted log\_revenue. The significant p-values (p < 0.05) explains that these values are statistically significant. "Everyone 10+," "Mature 17+," and "Teen" categories are estimated to contribute further predicted log revenues of 0.7946, 0.7875, and 0.7819, respectively, compared to "Everyone" category. This implies that targeting specific age groups can lead to a higher app revenue generation. From the marginsplot (fig 10.), it is clear that age groups such as Everyone 10+, Mature 17+, and Teen groups produce more revenue compared to everyone group, with predicted log\_revenue of 13.86669, 13.13.85957, 13.85395 and 13.07206 respectively. So company can focus mainly on Everyone 10+ and Teen groups for more revenue. Developing apps which are more interactive and educational with appealing visuals and user-friendly interfaces will be more used by pre-teen users of Everyone 10+ age group. When considering Teen group, apps related to games, social networking and photography can be given more importance.

A graph with a line and a point

Description automatically generated with medium confidence

**Fig 10: Marginsplot for Age Group and Log\_revenue**

4.2 Differential effect of app rating on revenue (logged) for different monetization strategies:

The baseline model has been changed to find the differential effect of rating on log\_revenue (Table 6. Differential effect of app rating on revenue (logged) for different monetization strategies ) for various monetization strategies.

The model has an R-squared value of 0.2530. The adjusted R-squared (24.34%) explains a reasonable fit. The model is statistically significant (F(40, 3101) = 26.26, p < 0.0001), depicting that the included variables together contribute to explaining the variance in log revenue.

The coefficients for log price (0.1491459) and rating (0.4888076) indicate their positive relationship with log\_revenue. For one-percent increase in log\_price, log revenue is increased by approximately 0.15%. Similarly, one-unit increase in app rating increases the revenue by 48.8%.

In the interaction model between rating and monetization strategies there is a negative interaction for paid;ads strategy with rating (-0.84). This indicates that the positive effect of a higher rating on revenue is diminished for apps using the paid; ads strategy. Conversely, the interaction term for paid;inapp is positive (0.46), proposing that a higher rating contributes more positively to revenue for apps having the paid;inapp monetization strategy.

The direct effects of monetization strategies indicate that paid;ads has a positive coefficient of 3.88, while paid;inapp has a negative coefficient of -0.74. This depicts that apps having the paid;ads strategy is more likely to have higher revenues compared to the apps utilizing paid;inapp strategy.

|  |  |
| --- | --- |
|  | (1) |
|  | Log of revenue |
| Log of Price | 0.149\*\*\* |
|  | (0.044) |
|  |  |
| App Rating | 0.489\*\*\* |
|  | (0.065) |
|  |  |
| paid | 0.000 |
|  | (.) |
|  |  |
| paid;ads | 3.879\* |
|  | (1.610) |
|  |  |
| paid;inapp | -0.744 |
|  | (0.785) |
|  |  |
| paid # App Rating | 0.000 |
|  | (.) |
|  |  |
| paid;ads # App Rating | -0.839\* |
|  | (0.391) |
|  |  |
| paid;inapp # App Rating | 0.460\* |
|  | (0.187) |
|  |  |
| Everyone | 0.000 |
|  | (.) |
|  |  |
| Everyone 10+ | 0.805\*\*\* |
|  | (0.148) |
|  |  |
| Mature 17+ | 0.767\*\*\* |
|  | (0.231) |
|  |  |
| Teen | 0.789\*\*\* |
|  | (0.119) |
|  |  |
| No. 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 |

Standard errors in parentheses

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

**Table 6. Differential effect of app rating on revenue (logged) for different monetization strategies**

From the above findings, the best monetization strategy for high-quality apps, which is indicated by higher app ratings, is paid;inapp. Inspite of having a negative main effect, the positive interaction with rating exceeds the direct negative effect. On the other hand, paid;ads seems to be less advantageous for high-quality apps, as its negative interaction with rating shrinks the positive impact of a higher rating on revenue. Fig 11. illustrates the visual representation of the same.

A graph with a line

Description automatically generated

**Fig 11: Result of interaction between Monetization strategy and app rating**

**5. Diagnostics and Robustness Analysis:**

5.1 Heteroskedasticity

Null Hypothesis : H0 =model is homoskedasticity

Alternate Hypothesis : Ha = model is not homoskedasticity

Heteroskedasticity is checked for the baseline model using 2 tests;

* White’s Test (imtest, white)

The results of White's test suggest a rejection of the null hypothesis, indicating the existence of heteroskedasticity in the model. The chi-square statistic of 340.68 having 227 degrees of freedom and a p-value of 0.0000 provides strong evidence against the assumption of equal variance of errors.

Cameron & Trivedi's decomposition test examines the sources of heteroskedasticity. The largest contribution is from the Heteroskedasticity term (340.68), whereas Skewness (52.81) and Kurtosis (0.04) make relatively smaller parts. The total chi-square statistic for the decomposition is 393.53 having 266 degrees of freedom and a p-value of 0.0000, recommending the presence of heteroskedasticity.

* Breusch–Pagan/Cook–Weisberg test (hettest)

The chi-square statistic of 24.67 with 1 degree of freedom and p-value of 0.0000 indicates strong proof to reject null hypothesis.

Fig 12: illustrates the distribution of heteroskedasticity in the baseline model.

A blue dot diagram with red line

Description automatically generated

**Fig 12: Heteroskedasticity od baseline model**

|  |  |  |
| --- | --- | --- |
|  | (1) | (2) |
|  | Log of revenue | Log of revenue |
| App Rating | 0.521\*\*\* | 0.521\*\*\* |
|  | (0.061) | (0.067) |
|  |  |  |
| Log of 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) |
|  |  |  |
| No. 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 |

Standard errors in parentheses

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

**Table 7. Comparison of baseline and robust model**

To conclude, the model shows the presence of heteroskedasticity, indicating that the variance of the errors is inconstant across the data. Addressing this issue may involve applying heteroskedasticity-robust standard errors.

After running and evaluating robust model, it is found that there is no change in any of the model measures (Table 7.), except the F-statistic. There is a slight increase in the F-statistic of robust model to 30.52 from 27.26 of the baseline model, which represents minute improvement in robust model when compared to base model.

5.2 Quadratic effect of log\_price

A new variable price\_sq has been generated which is the square of log\_price. This is to evaluate the results of quadratic effect of log\_price on log\_revenue in the model.

The new model with log\_price squared shows significant changes and improvements compared to the baseline model. The overall F-statistic remains almost the same as 27.26 (baseline) to 27.22 (quadratic), depicting that the extended model maintains a good fit. R-squared has slightly improved from 0.2503 to 0.2550, indicating that the quadratic model explains more variance in log\_revenue. The co-efficient of rating remains positive and significant at 0.05 significance level, describing that higher revenue is associated with higher ratings. There is a remarkable increase in the co-efficient value of log\_price, from 0.14 in baseline to 0.51 in quadratic model, indicating a strong positive relation between log\_price and log\_revenue. The co-efficient for log\_price squared is negative (-0.1359126), which implies a non-linear relationship between log\_price squared and log\_revenue. Monetization strategies and age target groups continue to display a similar pattern as in baseline model, with paid;inapp having the strongest and positive relation with log\_revenue.

A graph with a curve

Description automatically generated

**Fig 13. Illustrates the graphical representation of log\_price on log\_revenue for quadratic effect**

From fig 13., it is better to have the log\_price of app somewhere around the point 2, which can be considered as the optimal point, where the highest revenue generation take place. If log\_price goes below this point, revenue also goes low. On the other hand, if log\_price is higher than the optimal point, users do not prefer to install apps which have high price and ultimately, revenue will be low. Because of these reasons, it is best to have the optimal point around

Log\_price=2.

|  |  |
| --- | --- |
|  | (1) |
|  | Log of revenue |
| App Rating | 0.517\*\*\* |
|  | (0.061) |
|  |  |
| Log of Price | 0.511\*\*\* |
|  | (0.093) |
|  |  |
| log\_price squarred | -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) |
|  |  |
| No. 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 |

Standard errors in parentheses

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

**Table 8. Regression table for quadratic effect**

5.3 Endogeneity Problem

The baseline model can have potential endogeneity problems that can reduce reliability of estimated coefficients. Omitted variable bias is a concern, as crucial factors influencing app revenue and independent variables may be missing. Reverse causality, explaining that revenue influences independent variables, challenges the unidirectional causal assumption. Dummy variables, for monetization strategies, may become endogenous if correlated with unseen factors affecting revenue.

In our baseline model, only a few variables such as app price, number of languages, device type etc are considered for the baseline model. But there can be other variables which can be considered such as app size and app version where most users prefer to have apps with smaller size and good quality. Also, users often prefer apps with latest versions for improved features.

Panel data helps to include time-invariant features from the error, which helps to reduce the bias.

**Appendices:**

1. lab code

///Random checks on dataset

browse

describe

codebook age\_target--28 missing

codebook revenue

codebook monetization\_strategies

codebook main\_category

codebook num\_langs

///labelling

label variable product\_id "Product ID"

label variable name "App Name"

label variable developer "App Developer"

label variable version "App Version"

label variable release\_date "App release date"

label variable devices "Device Type"

label variable active "Is Active?"

label variable price "App Price"

label variable is\_paid "Is Paid?"

label variable size "App Size"

label variable rating "App Rating"

label variable num\_langs "No. of languages"

label variable main\_category "App Category"

label variable monetization\_strategies "Monetization Stategy"

label variable age\_target "Age Group"

label variable revenue "Total Revenue"

///generation of log variables

gen log\_revenue=log(revenue)

label variable log\_revenue "Log of revenue"

gen log\_price=log(price)

label variable log\_price "Log of Price"

gen sub\_category=.

replace sub\_category =1 if main\_category=="Games"

replace sub\_category =0 if main\_category!="Games"

label variable sub\_category "Games/Non-Games"

///dummy variables creation

encode monetization\_strategies, generate (monet\_strat\_dum)

encode age\_target, generate (age\_target\_dum)

encode main\_category, generate (main\_cat\_dum)

///descriptive statistics///

tabstat log\_price log\_revenue num\_langs rating size , statistics(count mean sd min max) by( sub\_category ) columns(statistics) longstub

ztest log\_revenue, by(sub\_category)

oneway log\_revenue main\_category

pwcorr log\_revenue log\_price size rating num\_langs, star(0.05)

///Explanatory analysis///

graph bar (mean) log\_revenue, over(monetization\_strategies, descending) blabel(bar) ytitle(mean Revenue (logged)) title(Monetization Strategy vs mean Revenue(logged))

graph pie log\_revenue, over(age\_target) sort pie(1, color(green)) pie(2, color(pink)) pie(3, color(yellow)) plabel(\_all percent) title(Age target vs revenue(logged)) legend(on)

graph box log\_revenue, over(devices) ytitle(Revenue(logged))title(Devices vs Revenue(logged))

graph bar (mean) log\_revenue, over(devices) over(sub\_category, label(labcolor("red"))) bar(1, fcolor(midgreen)) bar(4, fcolor(midgreen)) blabel(bar) ytitle(Mean Revenue (logged)) title(Revenue(logged) for Diff Devices across Sub-categories)

histogram log\_revenue, normal ytitle(Probability Density) xtitle(Revenue (Logged)) title(Density plot for Revenue (logged))

histogram rating, normal ytitle(Density) xtitle(Rating) title(Density Plot for Rating)

twoway (scatter log\_revenue log\_price) (lowess log\_revenue log\_price), ytitle(Revenue (logged)) xtitle(Price (logged)) title(Price - Revenue)

///Main Regression Analysis

reg log\_revenue rating log\_price i.monet\_strat\_dum i.age\_target\_dum num\_langs i.main\_cat\_dum

margins monet\_strat\_dum

marginsplot

margins age\_target\_dum

marginsplot

reg log\_revenue log\_price c.rating##i.monet\_strat\_dum i.age\_target\_dum num\_langs i.main\_cat\_dum

margins monet\_strat\_dum

marginsplot

///Diagnostic Analysis

///heteroskedasticity

reg log\_revenue rating log\_price i.monet\_strat\_dum i.age\_target\_dum num\_langs i.main\_cat\_dum

predict fitted

predict resid, residual

twoway (scatter resid fitted), yline(0)

drop fitted resid

imtest, white

hettest

reg log\_revenue rating log\_price i.monet\_strat\_dum i.age\_target\_dum num\_langs i.main\_cat\_dum ,vce(robust)

///Quadratic effect of log\_price

gen price\_sq= log\_price \* log\_price

label variable price\_sq "log\_price squarred"

reg log\_revenue rating log\_price price\_sq i.monet\_strat\_dum i.age\_target\_dum num\_langs i.main\_cat\_dum

twoway (scatter log\_rev log\_price ) (lowess log\_rev log\_price )

twoway (qfit log\_revenue log\_price )

///table export

ssc install estout, replace

cd "C:\Users\aj01421\Downloads"

 // summary statistics  
set more off  
eststo clear  
estpost tabstat log\_price log\_revenue num\_langs rating size , statistics(count mean sd min max) by( sub\_category ) columns(statistics) listwise  
eststo s1  
esttab s1 using SummaryStatOutput.rtf, cells("count mean min max sd") replace label

 //correlation output  
set more off  
eststo clear  
estpost pwcorr log\_revenue log\_price size rating num\_langs, matrix listwise  
eststo c1  
esttab c1 using corr\_table.rtf, replace label unstack not

// regression output model 1 (baseline model)  
set more off  
eststo clear  
reg log\_revenue rating log\_price i.monet\_strat\_dum i.age\_target\_dum num\_langs i.main\_cat\_dum  
eststo m1  
esttab m1 using reg\_model\_1\_table.rtf, replace ar2(3) b(3) se(3) r2(3) label compress

 // regression output model 2  
set more off  
eststo clear  
reg log\_revenue log\_price c.rating##i.monet\_strat\_dum i.age\_target\_dum num\_langs i.main\_cat\_dum

eststo m1  
esttab m1 using reg\_model\_2\_table.rtf, replace ar2(3) b(3) se(3) r2(3) label compress

 //regression output Model 1 and Model 3 comparison  
set more off  
eststo clear  
reg log\_revenue rating log\_price i.monet\_strat\_dum i.age\_target\_dum num\_langs i.main\_cat\_dum  
eststo m1  
reg log\_revenue rating log\_price i.monet\_strat\_dum i.age\_target\_dum num\_langs i.main\_cat\_dum ,vce(robust)  
eststo m3  
esttab m1 m3 using reg\_comp\_table.rtf, replace ar2(3) b(3) se(3) r2(3) label compress

// regression output model 4  
set more off  
eststo clear  
reg log\_revenue rating log\_price price\_sq i.monet\_strat\_dum i.age\_target\_dum num\_langs i.main\_cat\_dumeststo m1  
esttab m1 using reg\_model\_4\_table.rtf, replace ar2(3) b(3) se(3) r2(3) label compress