Analysis of reviews of Entertainment apps on Google play store

Qing Zhang, Suraj Kumar, Xin Zhang, Xinyue Shu, Yao Xiao

1 Introduction

Experiments were conducted as the part of research into Entertainment App from Google play. 111 Restricted and Unrestricted Entertainment Apps were used in the experiment and they each had their Ratings and Reviews. Association between reviews and ratings and the interactive effect due to restrictions are analyzed in the report. So the purpose of our report is to explore how app ratings affect app reviews.

This report focuses on a different analysis level through summaries, boxplots, and linear relations. Section 2 describes the exploratory analysis of link among all variables under consideration using a table of summaries,box plots, and scatter plot. Section 3 consists of linear model regression fitting and analysis of the result. Section 4 states the conclusion based on the overall analysis.

Note:The reviews are in millions, but the ratings are within 10, so their scales are completely different.So we did log transformation of reviews to reduce them down to an order similar to ratings.

2 Exploratory Data Analysis

Summary statistics of the rating of the entertainment app are presented in the following table for each Content rating separately.This table shows that there were approximately twice as many Restricted app in the sample (74 compared to 37).First, the mean rating of the restricted app was 4.12 compared to 4.16 for the unrestricted app.Secondly, we see that the middle 50% of restricted app is 4.2 between Q1 3.9 and Q3 4.3.We also note that the middle 50% of unrestricted app is 4.3 between Q1 4.0 and Q3 4.5.Thirdly,we can also see that the variability in the restricted app,as measured by the standard deviation of 0.26,the standard deviation of 0.43 for the unrestricted app.

Table 1: Summary statistics on Rating by Content Rating of 111 observations.

Content.Rating	Count	Mean	St.Dev	Min	Q1	Median	Q3	Max
Restricted	74	4.124324	0.2557882	3.4	3.9	4.2	4.3	4.6
Unrestricted	37	4.159459	0.4258713	3.0	4.0	4.3	4.5	4.7

Summary statistics of the reviews of the environment app are presented in the following table for each Content rating separately.This table shows that there were approximately twice as many Restricted app in the sample (74 compared to 37).First, the mean rating of the restricted app was 11.09 compared to 10.51 for the unrestricted app.Secondly, we see that the middle 50% of restricted app is 10.86 between Q1 10 and Q3 12.6.We also note that the middle 50% of unrestricted app is 10.4 between Q1 9.1 and Q3 12.11.Thirdly, we can also see that the variability in the restricted app,as measured by the standard deviation of 2.09, the standard deviation of 2.21 for the unrestricted app.

Let's compute the correlation coefficient between our outcome variable Reviews and our continuous explanatory variables Rating

Table 2: Summary statistics on log(Reviews) by Content Rating of 111 observations.

Content.Rating	Count	Mean	St.Dev	Min	Q1	Median	Q3	Max
Restricted	74	11.09	2.09	5.71	10.0	10.86	12.60	15.78
Unrestricted	37	10.51	2.21	6.20	9.1	10.40	12.11	14.42

```
# A tibble: 1 x 1
    cor
  <dbl>
1 0.187
```

We can see that the correlation between log(Reviews) and Rating is 0.1873251.There is only a weakly positive relationship between log(Reviews) and Rating.

We do here the box plot analysis

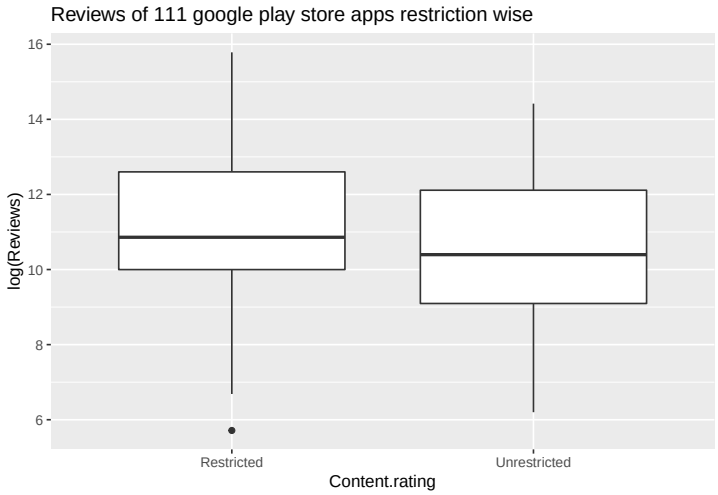


Figure 1: Rating by Content ratings

The boxplot shows that the restricted app having much reviews,in general,compared to the unrestricted app and that the reviews of the unrestricted were more widely distributed.There are also potentially one outlier which have unusually reviews.

We can now visualize our data by producing a scatterplot, where seeing as we have the categorical variable Content.Rating, we shall plot the points using different colors for each Content.Rating: There are very few restricted apps having less rating than 3.5 in our data set. There appears to be a positive relationship between log(Reviews) and Rating. Hence, log(Reviews) tends to increase with Rating. From the plotted regression lines, we can see that the lines have different slopes for restricted and unrestricted. That is, the associated effect of increasing rating appears to be more severe for restricted app than it does for unrestricted app, i.e. the reviews of restricted app raise faster with rating.

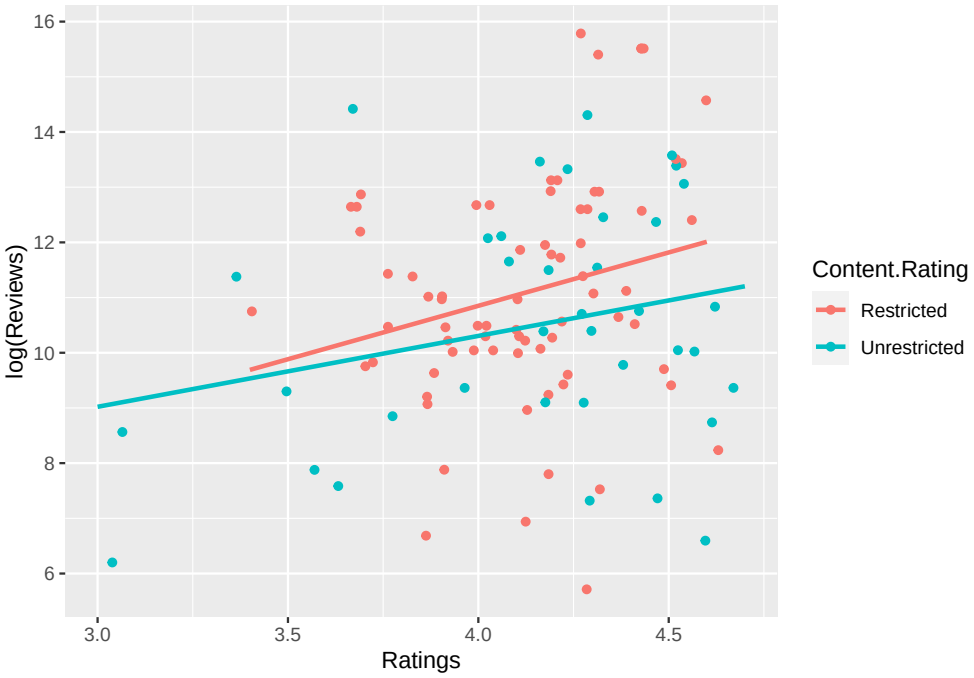


Figure 2: Relationship between Rating and Reviews. The best-fitting line has been superimposed.

3 Formal Data Analysis

We would like to start analysis by considering a full model. The full model considers the interaction between rating(continuous) and content.rating(categorical). The model can be expressed as:-

$$\widehat{\log(\text{Reviews})} = \hat{\alpha} + \hat{\beta}_{\text{Rating}} + \hat{\beta}_{\text{Content}} \cdot \mathbb{I}_{\text{Content}}(x) + \hat{\beta}_{\text{Rating * Content.rating}}$$

where:
* $\widehat{\log(\text{Reviews})}$ is the log transformation of reviews as a response variable;
* $\hat{\beta}_{\text{Rating}}$ is the cofficient for Rating varibale;
* $\hat{\alpha}$ is the intercept term;
* $\hat{\beta}_{\text{Rating * Content.rating}}$ is the interaction term coefficient;
* $\hat{\beta}_{\text{Content}}$ is the coefficient for Content.rating; and
* $\mathbb{I}_{\text{Content}}(x)$ is an indicator function such that

$$\mathbb{I}_{\text{Content}}(x) = \begin{cases} 1 & \text{if an entertainment app is unrestricted} \\ 0 & \text{Otherwise.} \end{cases}$$

Table 3: Estimates of the parameters from the fitted linear full model.

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	3.114	3.938	0.791	0.431	-4.693	10.920
Rating	1.934	0.953	2.030	0.045	0.045	3.823
Content.RatingUnrestricted	2.053	5.207	0.394	0.694	-8.270	12.376
Rating:Content.RatingUnrestricted	-0.649	1.254	-0.518	0.606	-3.135	1.836

From the table 3, we noticed that the coefficient $\hat{\beta}_{\text{Rating}}$ is positive. The value of $\hat{\beta}_{\text{Rating}}$ * Content.rating is less than 0,that means for a given rating, restricted apps will have more reviews. However, the confidence interval (-3.135,1.836) contains 0, therefore, the interaction term is insignificant. Likewise, $\hat{\beta}_{\text{Content}}$, which has confidence interval (-8.270, 12.376) contains 0, hence is insignificant too. We need to remove the two insignificant variables one by one. Let's fit a parallel model that ignores interaction term.

$$\log(\widehat{\text{Reviews}}) = \hat{\alpha} + \hat{\beta}_{\text{Rating}} + \hat{\beta}_{\text{Content}} \cdot \mathbb{I}_{\text{Content}}(x)$$

Table 4: Estimates of the parameters from the fitted linear parallel model.

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	4.661	2.557	1.822	0.071	-0.409	9.730
Rating	1.559	0.617	2.526	0.013	0.335	2.783
Content.RatingUnrestricted	-0.635	0.418	-1.518	0.132	-1.465	0.194

From the Table 4, $\hat{\beta}_{\text{Content}}$ has still a confidence interval(-1.465,0.194) which contains 0. Therefore, the factor term is insignificant. We work to further reduce the model to simple that only considers rating variable. lets analyse a simple model:-

$$\log(\widehat{\text{Reviews}}) = \hat{\alpha} + \hat{\beta}_{\text{Rating}}$$

Table 5: Estimates of the parameters from the fitted linear simple model.

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	4.65	2.573	1.807	0.073	-0.449	9.749
Rating	1.51	0.620	2.436	0.016	0.281	2.740

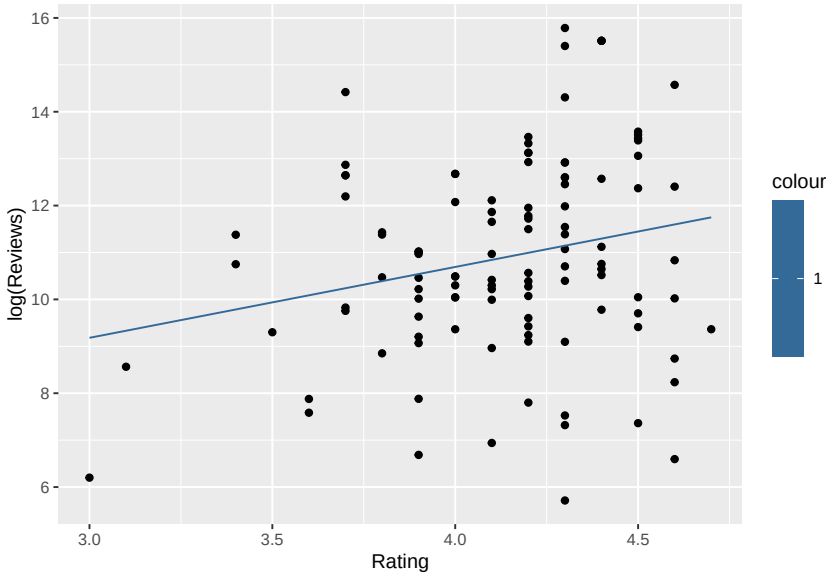
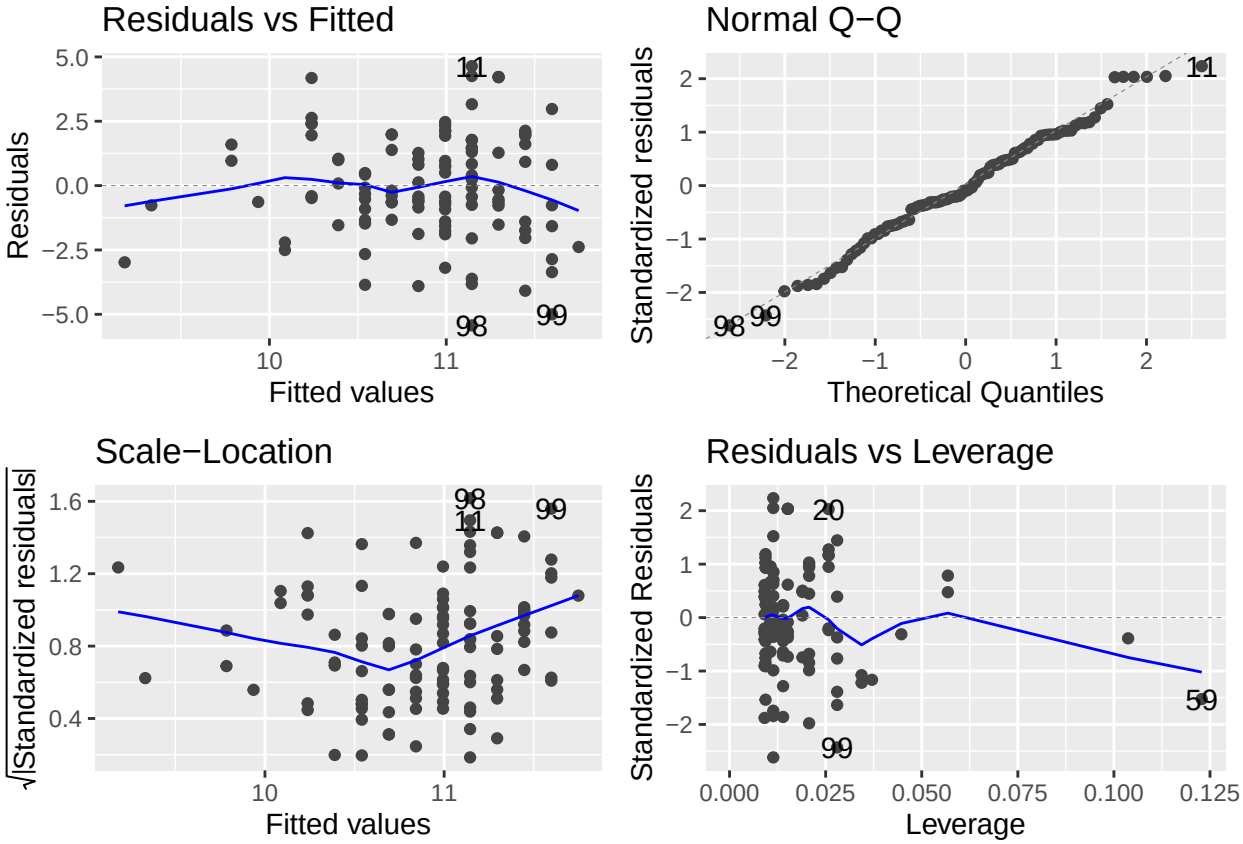


Figure 3: Simple linear fitting

Finally, we came to a model which has all parameters significant. The value of $\hat{\beta}_{\text{Rating}}$ is 1.51 and likely to lie in range between 0.281 and 2.740. The confidence interval doesn't contain 0. Log(reviews) shows a positive association with rating. For 1 unit increases in rating, log(reviews) would increase by 1.51 times. This is the right time to check if all standard assumptions hold for our final model.



we plot residuals vs fitted value and normal Q-Q plot to verify the standard assumptions of linear regression. In our case, there is no pattern observed in residual plot and normal Q-Q plot perfectly fits the diagonal line. Therefore, the residuals have mean 0, constant variance, and normal distribution. Moreover, we don't identify any outlier as no point has crossed the cook's distance line in residual vs leverage plot. Furthermore, we would like to check the variable selection method to consolidate our analysis.

Table 6: Model comparison values for different models.

Model	adj.r.squared	AIC	BIC
Interactive	0.05	483.80	497.35
Parallel	0.05	482.08	492.92
Simple	0.04	482.42	490.55

We noticed that the model3 has the lowest Bic value 490.55. The model3 has almost same Aic value of model2, but model3 has advantage of having less explanatory variable. Therefore, on the basis of selection criteria, we choose model3. We find our result coherent with the previous model analysis.

4 Conclusions

From our analysis, we can conclude that greater the rating, more would be the reviews. Reviews, however, don't depend upon the restrictions at all. Reviews will increase by a factor of 4.481689 for each unit increase in Rating and the increment is likely to be in the interval (1.32, 15.48). For an entertainment app ,having rating close to 3,has estimated review around (7708). The conclusion supports our initial impression.

5 Extention

We will consider more category of apps. We may include age-wise restrictions in the categorical variable. We may also include origin of the app country-wise as a part of analysis. Therefore, we can analyse how reviews are affected by country , age-restrictions, and genre.