Amazon Sales Data Review: Exmaining the effects of an item's discount percentage on the review star rating

Introduction

Sales is a fickle industry. The price an item is sold at, the amount of units sold, the specific color of the item sold, and more all depend on a variety of different metrics. One of these such metrics is review rating. The overall rating of an item can influence a consumer's decision to purchase an item, whether it's a comparison between two similar items or whether they need that specific item at all. Review scores are so important that in the restaurant business many owners when first starting out will actively reply to reviews left on rating websites such as Google or Yelp to apologize, thank, or sometimes contest a review left by a consumer because it can be vital to the success of building a reputation. With this in mind, I came across sales data scraped from Amazon.com's India locale to answer how an item's star rating is affected by the discount percentage of said item. This report aims to identify whether a causal relationship exists between star rating and discount percentage, examine the correlation between star rating and discount percentage accounting for the item category, review counts, and respective prices, and check for potential dependencies amongst review counts and item prices (discounted and actual). Initial assumptions and intuition after a cursory review of the data can lead us to believe that the larger the discount the potentially less favorable the star rating when accounting for review count, price difference, and the item category type.

Data

The data being analyzed is the Amazon Sales dataset taken from Kaggle.com. The dataset contains information scraped from India's Amazon.com marketplace in January of 2023 with 1465 total observations and each observation being the record of an item listed for sale. The dataset includes the following key variables examined:

- `category` - The category(s) that each particular product falls under. Categories fall under three major buckets, electronics, computers, and home/kitchen
- `discounted_price` - The discounted price of the product, if a discount is being offered. Currency is Indian Rupee, renamed to `dcntprice`
- `actual_price` - The non-discounted price of the product. Currency is Indian Rupee, renamed to `actprice`
- `discount_percentage` - The percent the listed product was discounted for, if a discount was available. Selected as the **explanatory variable**, renamed to `dcntpct`. Percentages listed were divided by 100 to turn them into decimals between 0 and 1

  `rating` - The average star rating of each product from 1 to 5. Selected as the **dependent variable**, renamed to `stars`
- `rating_count` - The total number of reviews for each product available, renamed to `reviewcnt`

The following dummy variables were created from the dataset for further analysis:

- `computers` - Denotes a product that falls under the Computers category with a 1, 0 if not
- `electronics` - Denotes a product within the Electronics category with a 1, 0 if not

The following variables were calculated:

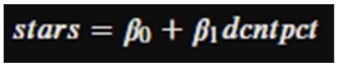
- `raw_dcnt` - Calculation of the price difference between the actual price and discounted price of a product
- `ln_reviewcnt` - The natural log of the `rating_count` column due to high variance of reviews per product
- `ln_raw_dcnt` - The natural log of the `raw_dcnt` variable due to high variance of price differences. Since the ln(0) is undefined, products that were not discounted were inputted as 0

Null values were examined in the dataset. There was 1 null value within `stars` and 2 within `reviewcnt`. For the purposes of this review, it could not be calculated what the null value for `stars` could represent so the value was dropped from the dataset. For `reviewcnt`, the review numbers were inputted as the mean value of reviews for the data, 18296 (rounded up) to not skew the data. A descriptive table can be seen in **Appendix A** of the numeric variables within the dataset.

Models

For this study I constructed multiple models for review of the data to test the hypothesis of star ratings being positively correlated with discount percentages increasing. Each model constructed is a linear regression with OLS using `stars` as the dependent variable and `dcntpct` as the explanatory variable. All regressions are performed with a covariance type of HC1 for a heteroskedasticity and robustness check. Regression results for all models discussed can be seen in **Appendices C1 and C2**.

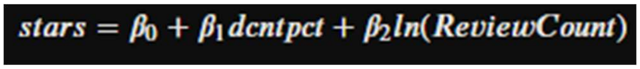
Model 1 - Star rating on discount percentage



$$stars = \beta_0 + \beta_1 dcntpct$$

From the results, an intercept of 4.197 and a coefficient of -.210 are seen, both statistically significant at p<.01 suggesting that the mean star rating when the discount percentage is 0 is 4.197 out of 5 and that for a 1% increase in the discount percentage it can be expected that the star rating will drop by .21 stars from the previous discount percentage. Two scatterplots were made using Seaborn to help illustrate this relationship. **Appendix B1** shows the scatterplot with the linear regression line through the data and **Appendix B2** shows the same plot but with a Loess line. The Loess line approximates the linear regression line however at the upper end of the data it curves up. However, for the vast majority of the data points available the slope of the Loess line is immaterial different from the slope of the line in **Appendix B1** and therefore a linear approximation of the relationship can be assumed.

Model 2 - Adding ln(Review Count) as a control



$$stars = \beta_0 + \beta_1 dcntpct + \beta_2 ln(ReviewCount)$$

The results are B_0 of 3.92, B_1 of -.178, and B_2 of .032, all significant at p<.01. This implies that when controlling for ln(Review Count) a 1% increase in discount percentage we can expect star rating to decrease by .178 and that for a 1% increase in review counts we can expect star rating to increase by .032.

Model 3 - Adding Computers dummy variable as an additional control

$$stars = \beta_0 + \beta_1 dcntpct + \beta_2 ln(ReviewCount) + \beta_3 computers$$

From these results B_1 decreases to -.220, indicating that when accounting for computer products a 1% increase in discount percentage more negatively impacts star rating. B_3 is .094 indicating that the expected star rating for a computer product is on average .094 higher than products in other categories. All values are significant at $p<.01$.

Model 4 - Replacing Computers with Electronics dummy variable

$$stars = \beta_0 + \beta_1 dcntpct + \beta_2 ln(ReviewCount) + \beta_3 electronics$$

Interestingly, B_1 increases to -.169 indicating that, while an increase to discount percentage still has a negative impact, it's less so when accounting for electronics instead of computer products specifically, significant at $p<.01$. B_3 is -.038 which is interpreted as electronics having a .038 lower star rating than other products comparatively, however this is significant at $p<.05$ instead of $p<.01$ meaning we have less confidence in this result being the true value.

Model 5 - Adding both dummy variables to compare against Home/Kitchen goods

$$stars = \beta_0 + \beta_1 dcntpct + \beta_2 ln(ReviewCount) + \beta_4 computers + \beta_5 electronics$$

B_1 decreases again to -.230 when accounting for all item categories. The coefficients on Computers (1.06) and Electronics (.020) are both higher when compared to home/kitchen products. The Electronics coefficient is not statistically significant so we cannot take confidence in this.

Model 6 - Adding in ln(Price Difference) as a control

$$stars = \beta_0 + \beta_1 dcntpct + \beta_2 ln(ReviewCount) + \beta_4 computers + \beta_5 electronics + \beta_6 ln(PriceDifference)$$

B_1 decreases further to -.274 ($p<.01$) and B_6 is at .011 significant at $p<.05$. The estimated coefficient for ln(Price Difference) indicates a slight increase to star rating for a 1% increase in price differences in items on sale of .011 stars however this is negligible.

Model 7 - Using ln(Price Difference) as lone control

$$stars = \beta_0 + \beta_1 dcntpct + \beta_2 ln(PriceDifference)$$

B_1 here is -.230 ($p<.01$) which is equivalent to model 5 and B_2 is .005 but not statistically significant. It can be inferred that the price difference does not have a significant impact on star rating on it's own.

Model 8 - Adding an interaction between ln(Price Difference) and ln(Review Count)

$$stars = \beta_0 + \beta_1 dcntpct + \beta_2 ln(ReviewCount) + \beta_3 ln(PriceDifference) + \beta_4 ln(PriceDifference) * ln(ReviewCount)$$

The resulting coefficients are as follows: B_0 3.758, B_1 -.187, B_2 .049, B_3 .025, and B_4 -.003 with betas 0, 1, and 2 being significant at p<.01 and betas 3 and 4 not being statistically significant. It's inferred from this that the interaction term does not have a significant impact on the model.

Summation of Findings

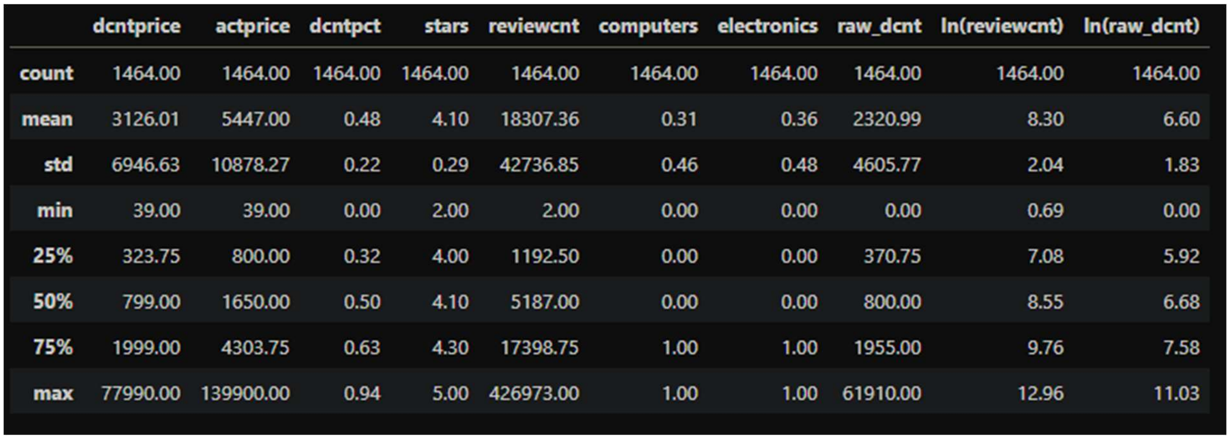
Amongst all models the analysis the of increases to discount percentage on star ratings were consistently negative. This can be seen as intuitive in that products that have a higher star rating do not necessarily require a discount in order to sell whereas a products with lower star ratings may need a discount to sell at a comparably. Interestingly, model 2 shows that when controlling solely for ln(Review Count) the impact of discount percentage on star rating isn't as strong. An interpretation of this could be that products' star ratings amongst similar review counts are not as impacted by increasing discount percentages compared to items that have larger varying review counts. A similar coefficient can be found in model 4 when accounting for products within the Computer category. Somewhat conversely, when accounting for Electronics rather than Computer products the coefficient *is* comparable to model 1's B_1 coefficient. It could be said that model 2 is moreso true for items in certain cateogries/types than others. When accounting for all control variables, the B_1 coefficient is the most negative out of any of the models at -.274 (p<.01). This implies that controlling for item categories and the change in raw price differences and not simply discount percentages along with the natural log of review counts that a 1% increase to discount percentage is more impactful negatively to star rating than when not controlling for other variables at all. The adjusted r^2 value for this model is the highest evidenced from the results at .094, however this is still a low value overall and indicates that the model is capturing an estimated 9.4% of the variance. When testing for an interaction between the natural logs of price differences and review count it was found that there was not a meaningful impact of these two variables on star rating or discount percentage as the results did not have statistical significance and coefficients fairly close to 0. When stepping back and checking for external validity, the data used for this analysis and the results from the models can be compared favorably with other markets and sales types. The products within the dataset are items that would be sold within many different markets globally as electronics and computers are similar if not exactly the same globally. Home/kitchen goods may vary more however a pot or pan in one country is comparable to a pot or pan in another country as well.

Conclusion

From the findings, it can be concluded that star rating is in fact negatively impacted by the discount percentage of the product, though not significantly so. When controlling for all types of variables, the negative impact remains, with varying strength. It cannot be said, however that the relationship is a causal one. The models used do show that there is a consistent negative correlation between star rating and discount percentages but more data and models would be needed in order to determine whether this is a causal relationship or not. In general, however, it can be recommended that if a seller has a product that they would like to sell, it would be wise for them to review their current star rating and set a discount percentage accordingly.

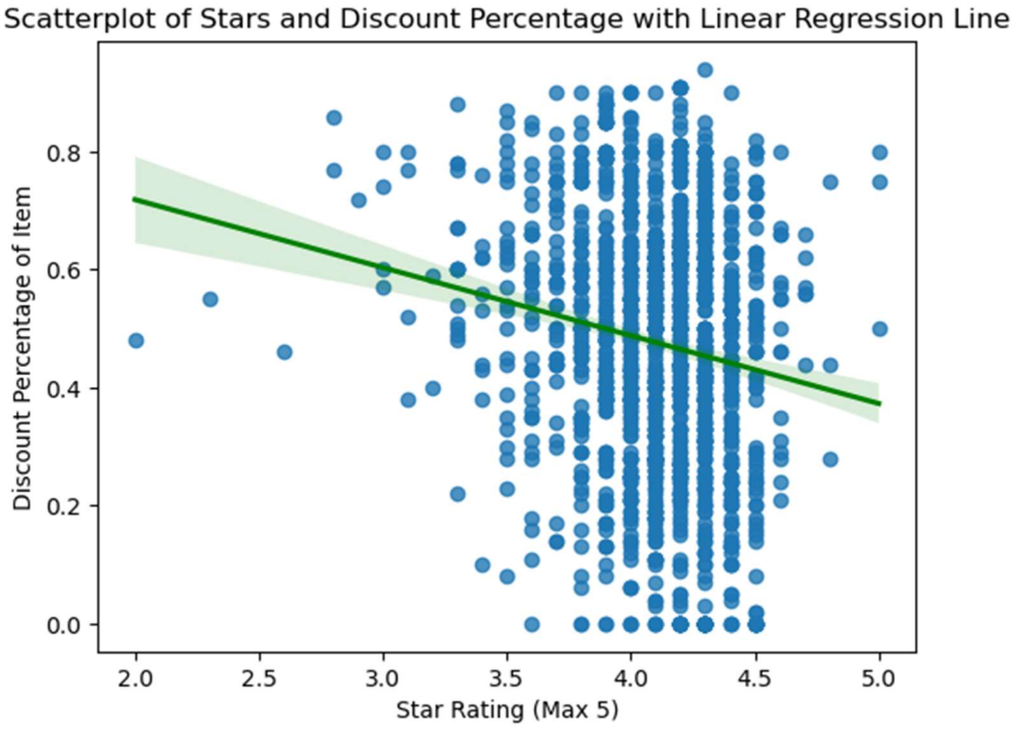
Appendices

Appendix A - Dataframe Descriptive Table

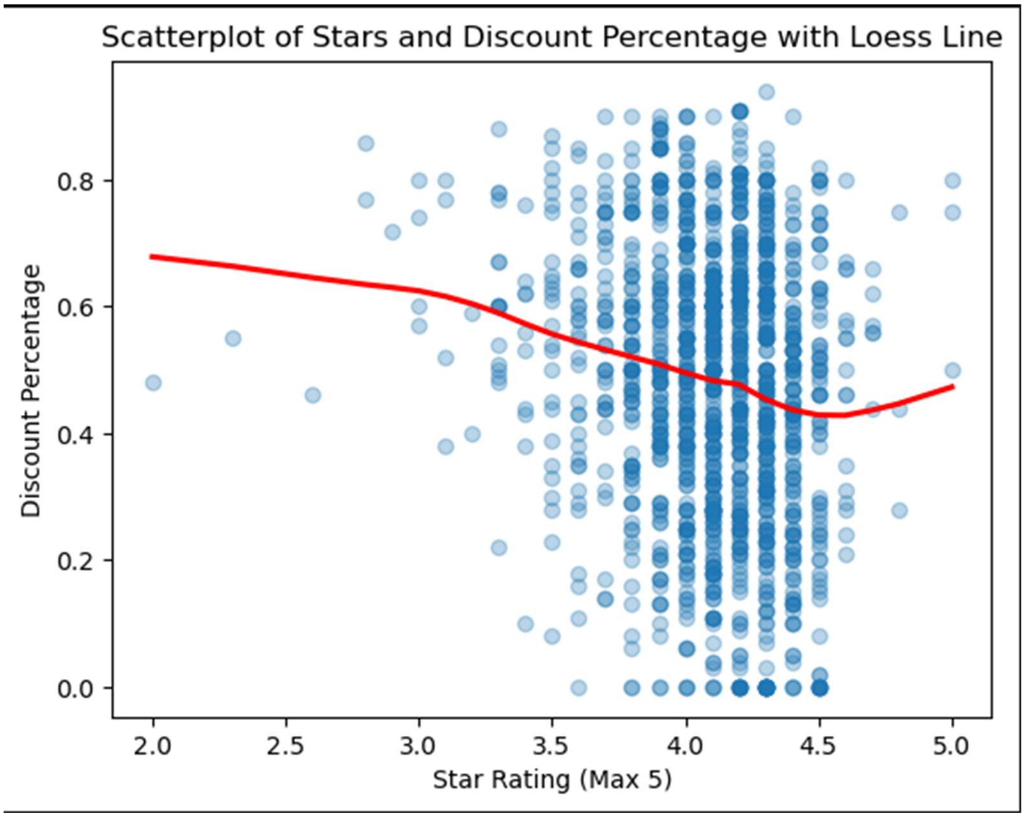


	dcntprice	actprice	dcntpct	stars	reviewcnt	computers	electronics	raw_dcnt	ln(reviewcnt)	ln(raw_dcnt)
count	1464.00	1464.00	1464.00	1464.00	1464.00	1464.00	1464.00	1464.00	1464.00	1464.00
mean	3126.01	5447.00	0.48	4.10	18307.36	0.31	0.36	2320.99	8.30	6.60
std	6946.63	10878.27	0.22	0.29	42736.85	0.46	0.48	4605.77	2.04	1.83
min	39.00	39.00	0.00	2.00	2.00	0.00	0.00	0.00	0.69	0.00
25%	323.75	800.00	0.32	4.00	1192.50	0.00	0.00	370.75	7.08	5.92
50%	799.00	1650.00	0.50	4.10	5187.00	0.00	0.00	800.00	8.55	6.68
75%	1999.00	4303.75	0.63	4.30	17398.75	1.00	1.00	1955.00	9.76	7.58
max	77990.00	139900.00	0.94	5.00	426973.00	1.00	1.00	61910.00	12.96	11.03

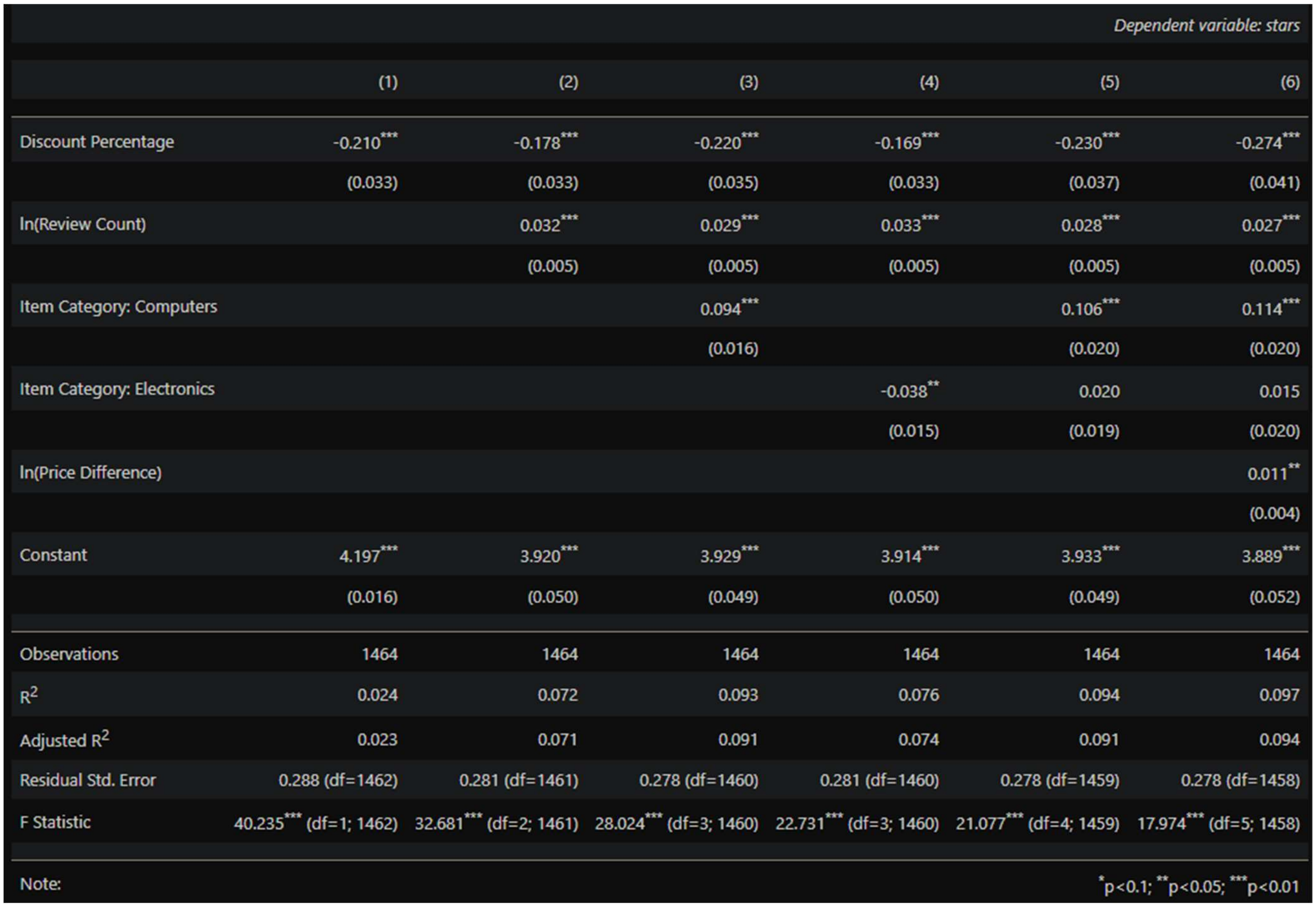
Appendix B1 - Scatterplot with Simple Regression Line



Appendix B2 - Scatterplot with Loess Line

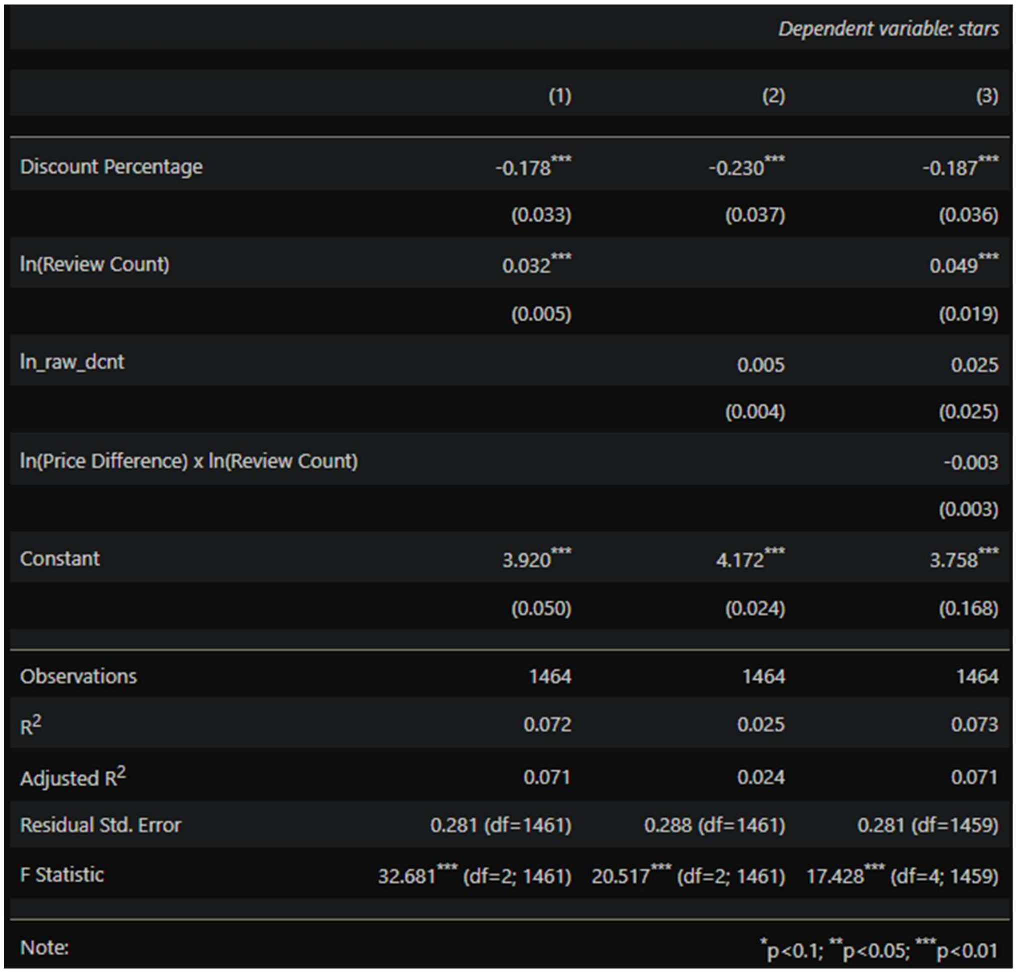


Appendix C1 - Regression results of first 6 Models compared



						Dependent variable: stars
	(1)	(2)	(3)	(4)	(5)	(6)
Discount Percentage	-0.210***	-0.178***	-0.220***	-0.169***	-0.230***	-0.274***
	(0.033)	(0.033)	(0.035)	(0.033)	(0.037)	(0.041)
ln(Review Count)		0.032***	0.029***	0.033***	0.028***	0.027***
		(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Item Category: Computers			0.094***		0.106***	0.114***
			(0.016)		(0.020)	(0.020)
Item Category: Electronics				-0.038**	0.020	0.015
				(0.015)	(0.019)	(0.020)
ln(Price Difference)						0.011**
						(0.004)
Constant	4.197***	3.920***	3.929***	3.914***	3.933***	3.889***
	(0.016)	(0.050)	(0.049)	(0.050)	(0.049)	(0.052)
Observations	1464	1464	1464	1464	1464	1464
R^2	0.024	0.072	0.093	0.076	0.094	0.097
Adjusted R^2	0.023	0.071	0.091	0.074	0.091	0.094
Residual Std. Error	0.288 (df=1462)	0.281 (df=1461)	0.278 (df=1460)	0.281 (df=1460)	0.278 (df=1459)	0.278 (df=1458)
F Statistic	40.235*** (df=1; 1462)	32.681*** (df=2; 1461)	28.024*** (df=3; 1460)	22.731*** (df=3; 1460)	21.077*** (df=4; 1459)	17.974*** (df=5; 1458)
Note:						*p<0.1; **p<0.05; ***p<0.01

Appendix C2 - Regression results of Models 2, 7 and 8



			Dependent variable: stars
	(1)	(2)	(3)
Discount Percentage	-0.178***	-0.230***	-0.187***
	(0.033)	(0.037)	(0.036)
ln(Review Count)	0.032***		0.049***
	(0.005)		(0.019)
ln_raw_dcnt		0.005	0.025
		(0.004)	(0.025)
ln(Price Difference) x ln(Review Count)			-0.003
			(0.003)
Constant	3.920***	4.172***	3.758***
	(0.050)	(0.024)	(0.168)
Observations	1464	1464	1464
R^2	0.072	0.025	0.073
Adjusted R^2	0.071	0.024	0.071
Residual Std. Error	0.281 (df=1461)	0.288 (df=1461)	0.281 (df=1459)
F Statistic	32.681*** (df=2; 1461)	20.517*** (df=2; 1461)	17.428*** (df=4; 1459)
Note:			*p<0.1; **p<0.05; ***p<0.01