# Lab2 : Estimating the Impact of Product Variations on Sephora Products Popularity

Anson Quon, Ishika Prashar, Naveen Sukumar

2023-12-08

## Introduction

The beauty and skincare industry has experienced remarkable growth over the years, and Sephora has created a lasting global presence for themselves in this industry. Given Sephora's expanding global footprint and its growing significance on online and mobile commerce, understanding the factors that influence virtual consumer behavior is imperative. In light of this goal, Sephora continues to enhance accessibility to its products by creating product variations that cater to a diverse range of consumers and their needs. To provide some context, variations can refer to features like the size of the product, or shade ranges of a foundation, for example.

While increased inclusivity is both a necessary and impactful way to cater to and bring in customers to Sephora, a data focused approach is needed to measure the impact of such variations on product popularity. Understanding how variations can influence the popularity and likeability of a product can help Sephora curate a more targeted range of product variations.

Therefore, this explanatory study hopes to estimate the impact of increased product variation on product popularity. Our primary research question is the following: *Does increasing the amount of variations of a product increase its popularity?* Our causal diagrams can be found in the appendix section.

## Data and Methodology

The data in this study comes from Kaggle, collected through a web scraper in March 2023 directly from the Sephora website. Each row in the data represents a specific product, uniquely indexed by its product id. It is important to note that product variations have their own row and in total there are about 8,000 products in the dataset. There are various features related to that product such as the name of the product, the number of times a product has been favorited, the number of variations, product category, and whether the product is online only or limited edition.

We use the variable loves_count to operationalize our Y concept of popularity. This variable measures the number of people who have marked a product as their favorite. We make the assumption that a more popular product will be favorited more often by consumers. To operationalize X, we use the child_count variable which refers to the number of available variations a product has. Our regression table also includes additional covariates - online_only and limited_edition both of which are binary variables. Online_only is 1 if the product is only available online, and 0 otherwise. Limited_edition is 1 if the product is designated as limited edition and 0 otherwise. Price_usd is the variable used to operationalize the cost of the product; it is continuous and numerical. Reviews is a variable used to operationalize the amount of feedback a product has received, also continuous and numerical. We also had an additional column primary_category which refers to the overarching category of the product such as skincare, makeup, hair, etc. We operationalize this variable by one hot encoding the most common categories of skincare, hair, makeup, fragrance, and bath and body (out of 9 unique categories) to serve as indicator variables and point estimates in our regression. Finally, we chose not to use other variables such as sephora_exclusive, brand_name, ingredients, and size since those variables had low correlations with the target variable loves_count according to a correlation plot during our data exploration. These variables also had a lot of null values, which could bias our results.

We first randomly split our data into an exploration and confirmation set. The exploration set is 30% of the dataset, and the confirmation set is 70% of the dataset. There was no need to remove observations from the overall dataset, as no missing values were present in our variables of interest. Using our exploration set, we first created a scatter plot of the relationship between child_count and loves_count and their respective histograms. We found that for both variables, the data was clustered primarily near 0 with some rare high outliers (right-skewed), all values were greater than or equal to zero, and the loves_count variable spanned multiple orders of magnitude (0-140,000). Given these findings, we decided to use a log transformation on both variables, and the scatter plot of the log transformed variables seemed more linear compared to the original plot. However, we did have 0 values in our dataset, and so in order to deal with this issue, we finalized our transformation to be of the form log(x+1) for both child_count and loves_count [1]. For child_count, we believed that adding 1 before log transformation does not hurt interpretability since a product with 0 variations still retains its default version, which we count as 1 variation. For loves_count, we believed that adding a small constant does not matter as much as obtaining a more linear scatterplot between our variables. Though we hypothesized that the variations are impactful for product favorites/popularity we do acknowledge potential confounders that can cause omitted variable bias if not included in our regression. Figure 1 further demonstrates this confounder with the variables online_only and limited_edition. We see that there is indeed a difference in product favorites based on these two indicator variables and it seems to affect both variations and product favorites. To account for this, we run a total of four regressions with increasing potential confounders (as described above). It is important to note that the price_usd and reviews variables were log transformed given they had values that were positive, and heavily right skewed.

Our first regression is of the form:

$$\log(\text{number of favorites} + 1) = \beta_0 + \beta_1 \cdot \log(\text{number of variations} + 1)$$

With the inclusions of confounders for additional regression equations.

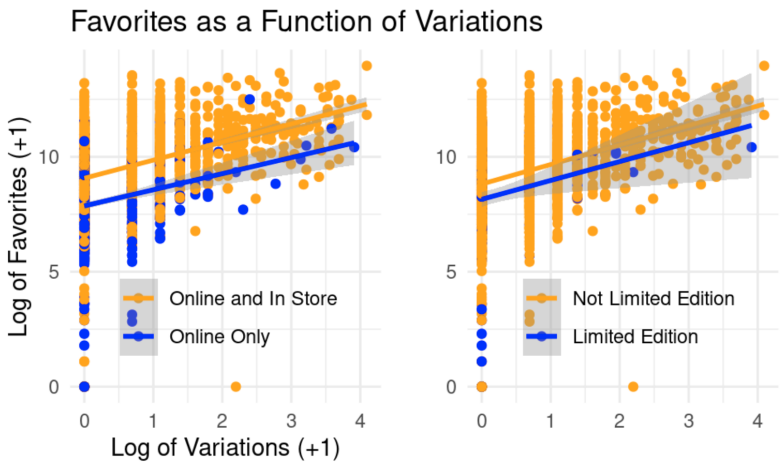



Figure 1: Scatter plot of variations vs favorites split by online only and limited edition

## Results

Table 1 shows the results of four representative regressions. Across all four models, the key coefficient on log(number of variations+1) is highly statistically significant. Point estimates range from 0.249 to 0.808. In model 1, we find that a 1% increase in the number of variations (+1) would result in a (approximately) .808% increase in the number of favorites (+1). In model 2, we find that holding online_only and limited_edition

[1] http://onbiostatistics.blogspot.com/2012/05/logx1-data-transformation.html - This blog post by a biostatistician discusses the common use of such a transformation. We further confirmed this approach with instructors Phillip and Majid who approved our decision.

constant, a 1% increase in the number of variations (+1) would result in a (approximately) .715% increase in the number of favorites (+1). In model 3, we find that holding online_only, limited_edition log(price_usd) and log(reviews) constant, a 1% increase in the number of variations (+1) would result in a (approximately) .352% increase in the number of favorites (+1). In model 4, we find that holding online_only , limited_edition log(price_usd), log(reviews) and the product category indicator variables constant, a 1% increase in the number of variations (+1) would result in a (approximately) .249% increase in the number of favorites (+1). (It is important to note that robust standard errors are not needed for our model since the residuals vs fitted plot shows that all four models have homoskedastic variance. The data initially did have very long tails but our log transformations helped mitigate this to be less extreme so we stick with classiical standard errors given they provide more accurate estimates.) In practical terms, it makes sense that an (%) increase in the variations of a product would result in an (%) increase in the number of favorites, as more variations would cater to a larger group of consumers. However, we do see that the effect of variations on number of favorites tends to decrease with the addition of covariates, in model 3 and 4 for example, we find that holding all else constant, a 1% increase in reviews would result in a (approximately) .491% and .489% increase in the number of favorites (+1), respectively. The % change by reviews seems to be much stronger than the % change by variations (+1). Perhaps models 1 and 2 had much higher coefficients on the number of variations (+1) because of omitted variable bias from reviews. Practically, it is possible that an increase in reviews may result in increased popularity as the product gains a larger consumer base, rather than variations, or perhaps reviews and variations are correlated positively and an uptake in reviews influences brands to increase variations. If we assume this, there seems to be a positive and away from zero direction of OVB.

Table1: Estimated Regressions

Results

| | Dependent variable: log(loves count + 1) | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| log (variations + 1) | 0.808*** | 0.715*** | 0.352*** | 0.249*** |
| | (0.028) | (0.027) | (0.020) | (0.022) |
| online only | | -1.137*** | -0.336*** | -0.275*** |
| | | (0.051) | (0.038) | (0.038) |
| limited edition | | -0.496*** | 0.717*** | 0.648*** |
| | | (0.082) | (0.063) | (0.063) |
| log price | | | -0.227*** | -0.161*** |
| | | | (0.020) | (0.022) |
| log reviews | | | 0.491*** | 0.489*** |
| | | | (0.009) | (0.009) |
| skincare | | | | 0.083 |
| | | | | (0.073) |
| hair | | | | -0.235*** |
| | | | | (0.076) |
| makeup | | | | 0.413*** |
| | | | | (0.073) |
| fragrance | | | | -0.106 |
| | | | | (0.080) |
| bath_body | | | | 0.226** |
| | | | | (0.094) |
| Constant | 8.740*** | 9.065*** | 7.667*** | 7.379*** |
| | (0.025) | (0.028) | (0.087) | (0.101) |
| Observations | 5,946 | 5,946 | 5,750 | 5,750 |
| R2 | 0.124 | 0.197 | 0.490 | 0.508 |
| Adjusted R2 | 0.124 | 0.196 | 0.490 | 0.507 |
| Residual Std. Error | 1.692 (df = 5944) | 1.620 (df = 5942) | 1.121 (df = 5744) | 1.101 (df = 5739) |
| F Statistic | 841.358*** (df = 1; 5944) | 485.500*** (df = 3; 5942) | 1,104.315*** (df = 5; 5744) | 593.134*** (df = 10; 5739) |
| Note: | | | | *p<0.1; **p<0.05; ***p<0.01 |

## Limitations

For large sample linear models such as ours, two assumptions must be made - IID data and a unique BLP exists. The products are identically distributed as they are collected from one point in time from a singular website. However, it is unclear if products would be independent. Given that many brands have multiple different products available on Sephora, there could be clustering by brand which we are not able to account for in our model. Perhaps some brands tend to be more popular than others or luxury vs affordable brands have different clusters, and there is also competition between the brands which can influence the increase or

decrease in variations or popularity. Given the potential of clustering by brands, we essentially have less data than we think we do, and thus our calculated standard errors may be narrower than they should be if IID were met. Moving on to the best linear predictor assumption, the potential for non-finite variance is possible because the variables loves_count, child_count, price_usd, and reviews were heavily right skewed both pre and post log transformations. We do find that there is no perfect collinearity between our variables because no variables were dropped from our regressions. Even though our sample will always provide a finite variance, the long tails suggest that our underlying distribution could have non-finite variance and thus there is a possibility that no unique BLP may exist for our distribution.

There are also additional omitted variables that may be biasing estimates. Within the dataset, there are variables such as sephora_exclusive, out_of_stock, and ingredients which could provide information regarding the quality of a product. This information could be used to infer upon a product's popularity. In addition, the model also does not take into account external factors that the dataset itself does not contain, factors such as current trends on social media, the marketing campaigns of certain brands or products, the presence and effects of promotions/discounts on certain brands or products, and etc. These external factors also contain important information that can be heavily correlated with the popularity of a product. For example, a viral influencer may be promoting a certain product, causing more people to favorite the product. The influencer's promotion could also lead to more variations such influencer's tend to have audiences with specific preferences. This in turn would lead to a positive omitted variable bias away from 0 for the variable number of variations.

Finally, another limitation to consider is reverse causality. While our analysis indicates that increased product variations can lead to an increased number of favorites, we can not say for certain that this is the only causal relationship, perhaps the causality exists in the opposite direction as well. Seeing a product have more favorites than another could cause brands to create more variations of this product to capitalize off its success. If this is the case, and the impact of variations and the number of favorites is positive, and vice versa, a positive feedback loop is created resulting in a positive bias in our current estimates.

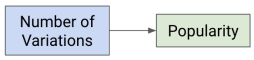
## Conclusions

This study aimed to analyze whether the variations of products at Sephora can influence their popularity. Based on our analysis, increasing the variations of a product has a positive influence on its popularity, operationalized by the number of times the product has been favorited. We believe that this relationship is valid in the real world since more variations of a product would cater to a larger population. However, we note that adding more covariates such as online_only, limited_edition, price_usd, and reviews decreases the effect of the number of product variations on the number of favorites given to the product. Nevertheless, the estimates are consistently highly statistically significant. Therefore, we conclude that increasing the number of variations of a product, does increase its popularity.

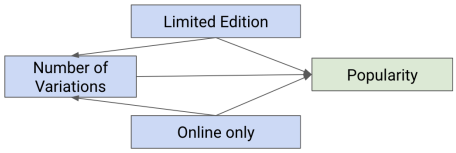
In the future, new data may be collected to specify various types of popularity metrics, such as by total sales, or number of social media mentions. This research can be applied to other companies and products to evaluate whether the same effects are observed or not. Perhaps our findings are generalizable, or they may be unique to Sephora specifically.

# Appendix

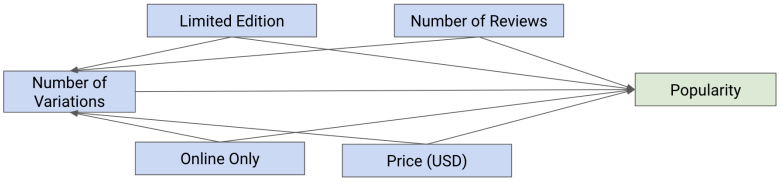
Model 1:




Model 2:




Model 3:




Model 4: