Location, Location, Location: How Product Placement Impacts Clicks

DATASCI 203: Lab 2

Sophie Chance, Amy Zhang, Maureen Fromuth

Introduction

An effective online store has become critical with the growing relevance of e-commerce. With limited consumer attention spans and webpage real estate, it is important for e-commerce companies to understand the impact of various factors of product advertisement. As an online retail strategy analytics team, we believe that it is important for retail owners, such as our client maternity store, to understand the impact of location placement of a product on clicks generated. This information can be leveraged by retail owners and software teams to utilize webpage locations to increase clicks for priority products.

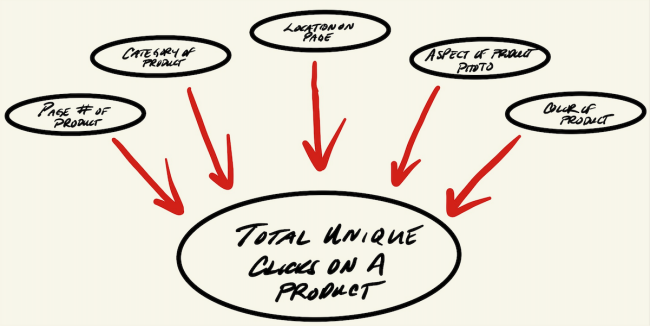


Figure 1: Causal Pathway for Clickstream Study

This study will use clickstream data to evaluate whether the location of a product on a webpage has any statistical significance in altering the unique-per-session clicks the product receives off of a baseline location and page number. Applying a set of regression models, our team has determined the statistical significance and estimated the relative changes in number of clicks that result from the placement of an item on a webpage.

Given the languages of our client's customers are all left-to-right (LTR), the practical hypothesis of this test is that placing products in the top left position will result in greater number of total unique clicks per session. Statistically, the null hypothesis for this study is that the location of the product has no impact on the number of unique session clicks the product receives.

Description of Data

We will be leveraging a dataset containing observational information on clickstream from an online store offering clothing for pregnant women. The data is from five months of 2008. While this data is aged and e-commerce user experience has greatly improved in recent years, we believe that the general learnings around product placement and clicks generated will still hold.

The data represents a total of 165,474 clicks on 218 products. Each click identifies the date of the click and online session ID. The data also provides product details, including the product category, the color, and the price in dollars. It also identifies the location of the product photo on the webpage, page number of the product, and whether the product photo is profile or face on. The product and website layout are static over the course of the data collection. As such, we selected the individual product as our unit of observation.

Operationalization of Key Concepts

While we considered other variables for the treatment variable, such as page number or price, we believed that the impact of the specific location on the page was less studied in this field, and therefore more valuable for our client to understand. To analyze the number of clicks per product, we grouped the original dataset by product. We chose the variable 'location of product' to represent the treatment variable of location on the webpage. For the output variable, we selected 'total clicks per product' to represent the total unique session clicks per product.

The selected variables match the concepts well, but some transformations were necessary to minimize the gap between the conceptual and operational definitions. We renamed columns to enable readability and we also transformed numerical values to the named value (e.g. we transformed the value 1 in 'location on page' to 'top-left').

X or Y	Concept	Actual Feature Used
X	Location of Product	Location of Product (e.g. 'top-left')
Y	Total Number of Clicks per Product	Sum of clicks (rows) by product, counting 1 click per session

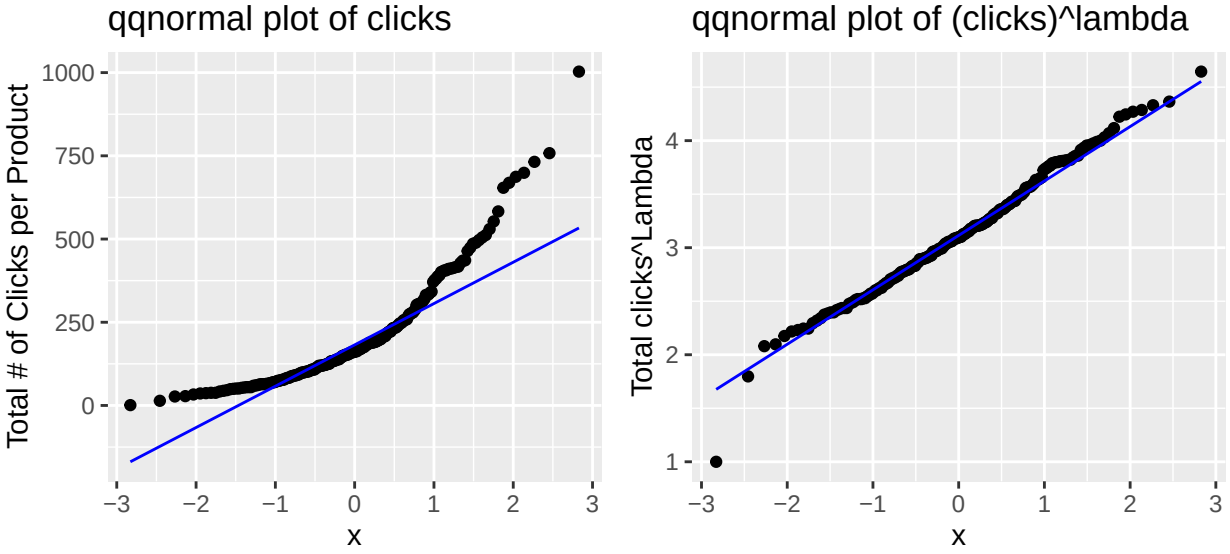
We also eliminated portions of our data prior to modeling to limit dependence and influential outliers. For 'total clicks per product', we aggregated the sum of clicks for each product, but only counted one click per product per session. This allowed us to mitigate the impact of certain sessions where users clicked repeatedly on the same product, but there still exists a gap to the ideal operationalization as the dataset does not allow us to ensure independence across sessions. We also wanted to control for product popularity, namely products that were particularly faddish and had a notably high number of clicks. Recognizing that outliers are particularly impactful for linear models with categorical variables, we removed outlier products from the exploratory dataset based on Cook's distance (*NOTE* see next section for details on our exploration vs. confirmation dataset).

Dataset	Element Changed	Amount Removed
Original Dataset (rows = clicks)	Repeat clicks counted on the same product within the same session	13,065 clicks removed
Exploratory Dataset, Grouped by Product (rows = products)	Outlier Products based on Total Unique Session Clicks per Product	7 products removed

Explaining Key Modeling Decisions

We split our data into exploration (30%) and confirmation (70%) sets at the individual session level.

In our exploration dataset, we first evaluated the normality of our output variable, total number of unique clicks per product. A histogram of the distribution revealed a significant right skew to the total number of clicks. Using the Box Cox test, we concluded that raising our output variable to a power of 0.2222222 (lambda) would result in greater normality. This conclusion was further supported by a visual comparison of the qqnorm plots and the results of the Shapiro-Wilks Normality test for the transformed variable.



Of note, we reran the Box Cox test following the removal of outliers. This resulted in a decrease in the Lambda value to 0.0606061, which was validated by a second Shapiro-Wilks Normality test. For the final model, we selected 0.0606061 for our final transformation value for total unique session clicks.

We also evaluated which variables to incorporate into the model. Based on our analysis, we incorporated the product location on page, product page number, and product category. To determine the appropriate variables, we evaluated the impact of omitted bias to the coefficients of the location variable. We balanced the assessed impacts of omitted variable bias with the need to prevent model overfitting. We omitted the price of the product as the price is not displayed on the page, so the consumer decision to click would be not impacted by the price of the product. We omitted the aspect of the photo as there is a negligible difference between the two options, and was therefore not beneficial to include in the model. Lastly, we omitted color as there were 10+ color options, and including this variable would greatly increase the degrees of freedom for the model and contribute to potential overfitting.

Finally, we selected the base for our categorical models based on our null hypothesis. The baseline position for our model will be for a sale product that is located in the top-left corner of the page and on the first page of the website.

Results

Table 3: Estimated Regressions

	Output Variable: Unique Clicks per Product		
	(1)	(2)	(3)
Pos: Bottom Left	−0.005 (0.02)	−0.01 (0.02)	−0.01 (0.02)
Pos: Bottom Middle	0.01 (0.02)	0.002 (0.01)	−0.002 (0.02)
Pos: Bottom Right	−0.01 (0.02)	−0.02 (0.02)	−0.03 (0.02)
Pos: Top Middle	−0.0003 (0.02)	−0.003 (0.02)	−0.02 (0.02)
Pos: Top Right	−0.03 (0.02)	−0.03 (0.02)	−0.03 (0.02)
Page 2		−0.11** (0.04)	−0.12** (0.04)
Page 3		−0.09*** (0.02)	−0.09*** (0.02)
Page 4		−0.08*** (0.01)	−0.07*** (0.01)
Page 5		−0.07*** (0.01)	−0.07*** (0.01)
Category: Blouse		0.04*** (0.01)	0.03* (0.02)
Category: Skirt		0.06** (0.02)	0.07* (0.03)
Category: Trousers		0.07*** (0.01)	0.08*** (0.02)
Price			−0.001 (0.001)
Base: Top Left, Page 1, Sale	1.43*** (0.01)	1.45*** (0.01)	1.49*** (0.03)
Product Color			✓
Model Aspect			✓
Price Higher than Average			✓
Observations	218	218	218
R^2	0.02	0.36	0.42
Residual Std. Error	0.09 (df = 212)	0.07 (df = 205)	0.07 (df = 189)
Note:		HC_1 robust standard errors in parentheses.	

Table 1 shows the detailed findings of the 3 regressions run. These models did not find significant evidence to reject the null hypothesis that the location of the product does not affect the total number of unique clicks. While this is true regarding overall product placement, there was a marginally significant impact for the top right location. Model 2 reveals that if a product is placed in the top right location, the total number of clicks is reduced by 3%, and should be avoided by our client for priority products where it is important to increase distinct clicks.

Of the other variables tested, the page number and product category had highly significant p-values, reflecting a significant impact on the number of unique clicks that a product would receive. While the product category is not necessarily an actionable insight, it would be unwise for our client to ignore the significance of the page number on the number of clicks. By placing a product on the second page, the product saw a reduced number of clicks by 7.3%, which is an intuitive finding. Our client should use this finding to place priority products on earlier pages.

Surprisingly, the base case of an item being on sale did not yield more clicks on the product. In the most extreme example, there is a 7.16% increase in the number of clicks on trousers over sale items. This could be due to the lack of price details on the page, or that already poor performing products are placed on sale.

Limitations

This model is based on a large sample of data, but we evaluated the assumptions for Classic Linear Models to identify potential areas for improvement. Below lists those assumptions that were potentially violated in our final model.

IID Data *evidence of geographic and session-based dependence* The dataset does not provide customer information for each click. We partially accounted for independence concerns by deduping clicks that occur in the same session. However, we cannot guarantee that the observations are therefore independent as there is no tracking possibility across sessions. For example, there may remain geographic clustering with specific cities or regions having unique click trends. This clustering would be particularly an issue for a model that includes other covariates such as color, which we did not include in our final model selection.

Linear Conditional Expectation *evidence of a non-linear relationship* A graph of the residuals v fitted values initially appears to indicate a random and consistent spread of data points around 0 on the x-axis. A histogram plot of residuals, however, highlights a significant left skew in the distribution. Given the spread of residuals was only 0.5, it is hard to assess if the linear conditional expectation assumption was fully violated.

Normally distributed errors *evidence of non-normality in residuals* We performed several tests to include the Shapiro-Wilk normality test, which resulted in a p-value less than 0.05. This resulted in a rejection of the null hypothesis of normality in residuals. Similarly, further evaluation of the distribution of the residuals demonstrated in a left skew and a significantly high kurtosis at 16 signifying a Leptokurtic distribution.

For structural limitations, our model may be biased by omitted variables. This dataset does not provide the products' popularity or dates of availability. For popularity, trendy items may see higher clicks regardless of its placement. We expect a positive correlation between product popularity and total clicks, but no relationship between product placement and popularity. Therefore, we expect a negative omitted variable bias on the key variables. The main effect is therefore driven towards zero, making our hypothesis tests underconfident. This is similar for the amount of time that a product is available. It is unclear how much time each product is listed on the website, resulting in a positively correlated bias with total clicks. However, we do not believe this OVB calls the results of this study into question, as we have accounted for outliers that may be unduly influenced by variables such as popularity.

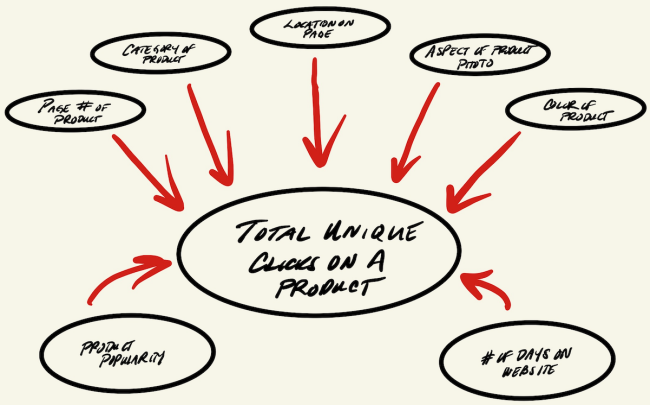


Figure 2: Updated Causal Pathway for Clickstream Study

Our team also evaluated reverse causality and outcome variables on the right hand side. We did not find any evidence of RHS outcome variables but reverse causality may be present in our product category, specifically with sales. With a lack of clicks for a product, the client may have moved it from its original product category to sales. Such a move would likely present a lower coefficient for sales than what is true.

Conclusion

This study found that product location on page had no statistically significant impact to total clicks on a product. Rather, the page of the product was statistically relevant, with a negative correlation between page number and number of clicks. The category of the product was also significant, and sales products saw decreased clicks. The client should leverage these findings as they consider the page of the product as an actionable insight when determining ordering of priority products, and pay less attention to the specific location on the page.

For future research, it would be helpful to design an experiment or leverage a new dataset with more recent data, which should give a larger sample size and depth of data that will allow the research team greater flexibility in model building. For example, it would be helpful to have a dataset where the same product exists in multiple locations on a webpage to allow researchers to better control for product feature covariates. It would also be helpful to have additional data on the potential omitted variables mentioned, including popularity of products and amount of time available on the website. Lastly, it would be helpful to have a customer ID included in this new dataset to augment our model based on repeat clickers and even potentially customer demographics.

Sources

1. https://www.slideserve.com/kamin/mariusz-apczy-ski-department-of-marketing-research-cracow-university-of-economics-cracow-poland

2. https://www.kaggle.com/code/ahmedsaed26/e-shop-clothing-eda-classification-and-regression

3. https://archive.ics.uci.edu/dataset/553/clickstream+data+for+online+shopping

4. https://library.virginia.edu/data/articles/interpreting-log-transformations-in-a-linear-model#:~:text=Only%20independent%2Fpredictor%20variable(s)%20is%20log%2Dtransformed.&text=This%20tells%20us%20th

5. http://www.sthda.com/english/articles/40-regression-analysis/163-regression-with-categorical-variables-dummy-coding-essentials-in-r/

6. https://www.youtube.com/watch?v=-1ERHZ5Hnx8

7. https://stats.oarc.ucla.edu/other/mult-pkg/whatstat/

8. https://stackoverflow.com/questions/33397689/multi-collinearity-for-categorical-variables

9. https://blogs.ubc.ca/datawithstata/home-page/regression/ordinary-least-square/#:~:text=Including%20as%20many%20d

10. https://r-coder.com/box-cox-transformation-r/#google_vignette

11. https://faculty.nps.edu/rbassett/_book/regression-with-categorical-variables.html

12. https://universeofdatascience.com/box-cox-transformation-for-normalizing-a-non-normal-variable-in-r/