Location, Location, Location … How Product Placement Impacts Clicks

Lab 2 Proposal

DATASCI 203, Fall 2023

Sophie Chance, Amy Zhang, Marty Fromuth

1. What is the research question? Specifically what is the X concept and what is the Y concept?

   Question - Holding all else equal, how does the position of an item on the page (X) affect how many clicks a product receives (Y)?

   Treatment Variable (X) - location of item on the page

   Performance Variable (Y) - total number of clicks on a product

   Potential confounders or covariates that may impact the output of the regression include color, category of product, price of product, model photography, and page number of the product. Given each of the products are fixed and only in one location, we will compare how the coefficient of the location variable changes as we incorporate additional covariates into our model.

   We will pay particular attention to the general popularity of the product, product rating, and product photo quality to assess if there are any omitted variable bias (OVB) or reverse causality. Of note, since the data does now allow us to determine if each click is from a unique/new viewer, we will also evaluate how the number of repeat viewers for a product would impact the coefficient of product location (i.e. if one user is responsible for 20% of the clicks, does that impact the true impact of location).

2. Who is the actor who can change your X concept?

   Web developer of this online store can change the positioning of items on the page.

3. Who is the audience who would care about changes in the Y concept?

   Under the assumption that more clicks increases the potential for more purchases, the E-commerce store will be interested in the results of this study. They will be able to strategically place products on the web page based on these results in an attempt to get more clicks on certain products.

4. What is the data source? What variables will you use to operationalize X and Y?

   Data Source -
   The dataset contains information on clickstream from an online store offering clothing for pregnant women. Data are from five months of 2008 and represents clicks on one

of 217 different products. These products fall in 4 product categories, have a price in U.S. dollars, a specific color, page number that the photo of a product is located, location of the photo for the product on the webpage, and whether the photo for that product is either profile or face on.

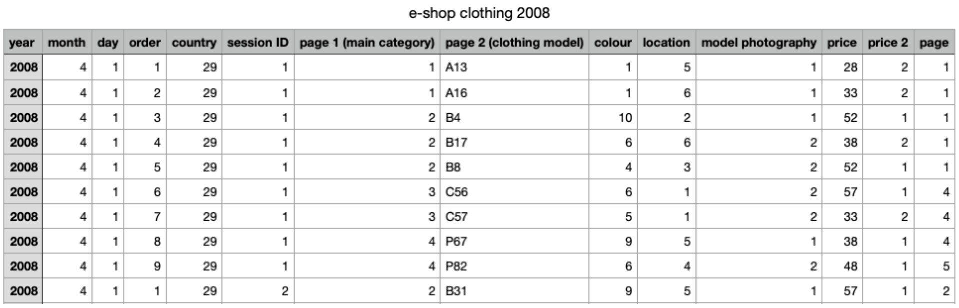
Citation of Data -
£apczyÒski M., Bia≥owπs S. (2013) Discovering Patterns of Users' Behaviour in an E-shop - Comparison of Consumer Buying Behaviours in Poland and Other European Countries, ìStudia Ekonomiczneî, nr 151, ìLa sociÈtÈ de l'information : perspective europÈenne et globale : les usages et les risques d'Internet pour les citoyens et les consommateursî, p. 144-153.
https://www.kaggle.com/datasets/tunguz/clickstream-data-for-online-shopping

Variables -
X: 'location of product'
*** Will be based off of the 'location' variable in the current dataset; there are 6 options for the locations - top left, top in the middle, top right, bottom left, bottom in the middle, bottom right*
Y: 'total clicks per product'
*** Will be based off of the count of unique clicks on each product using the 'index' of that click.*

5. What is the unit of observation? That is, does each row of the data represent a person, a review, a hotel stay, or something else?

For this analysis, we will transform the original data from rows associated with a click on a product to a new dataset associated with the number of clicks per product.

Each line of the original data represents a click for a viewer on a unique online 'session'. There are 165,474 different clicks that are associated with one of 24,026 different online 'sessions'. For each click, features of the data include, among others, product category, location of the photo on the page, country of origin of the IP address and product price in US dollars.



e-shop clothing 2008

year	month	day	order	country	session ID	page 1 (main category)	page 2 (clothing model)	colour	location	model photography	price	price 2	page
2008	4	1	1	29	1	1	A13	1	5	1	28	2	1
2008	4	1	2	29	1	1	A16	1	6	1	33	2	1
2008	4	1	3	29	1	2	B4	10	2	1	52	1	1
2008	4	1	4	29	1	2	B17	6	6	2	38	2	1
2008	4	1	5	29	1	2	B8	4	3	2	52	1	1
2008	4	1	6	29	1	3	C56	6	1	2	57	1	4
2008	4	1	7	29	1	3	C57	5	1	2	33	2	4
2008	4	1	8	29	1	4	P67	9	5	1	38	1	4
2008	4	1	9	29	1	4	P82	6	4	2	48	1	5
2008	4	1	1	29	2	2	B31	9	5	1	57	1	2

For this study, we will transform the original data by doing a group_by() function on the data and grouping by product with features of the product (price, relative price, color, product category) as well as number of clicks for all sessions and number of clicks for all sessions per each of the location options. As such, for the purposes of this study we will evaluate clicks per individual products as our observation.

To offer some control on price, we will only evaluate products whose price is not above the average price for that product type/category. These controls will result in evaluation of 109 resulting products.

Additionally, we will attempt to control for the users that may read right to left by only including data with IP addresses from countries with primarily languages in Latin, Modern Greek, Cyrillic, Indic and Southeast Asian scripts.

Finally, to attempt to mitigate challenges of independence between sessions (i.e. one user may be responsible for multiple sessions, and may be influenced on what to click based on previous sessions), we will conduct a random sampling of the original dataset and then group by product as discussed above.