Lab 2: What Makes a Product Successful? Fall 2021

w203: Statistics for Data Science

November 1, 2021

5. Limitations of your Model

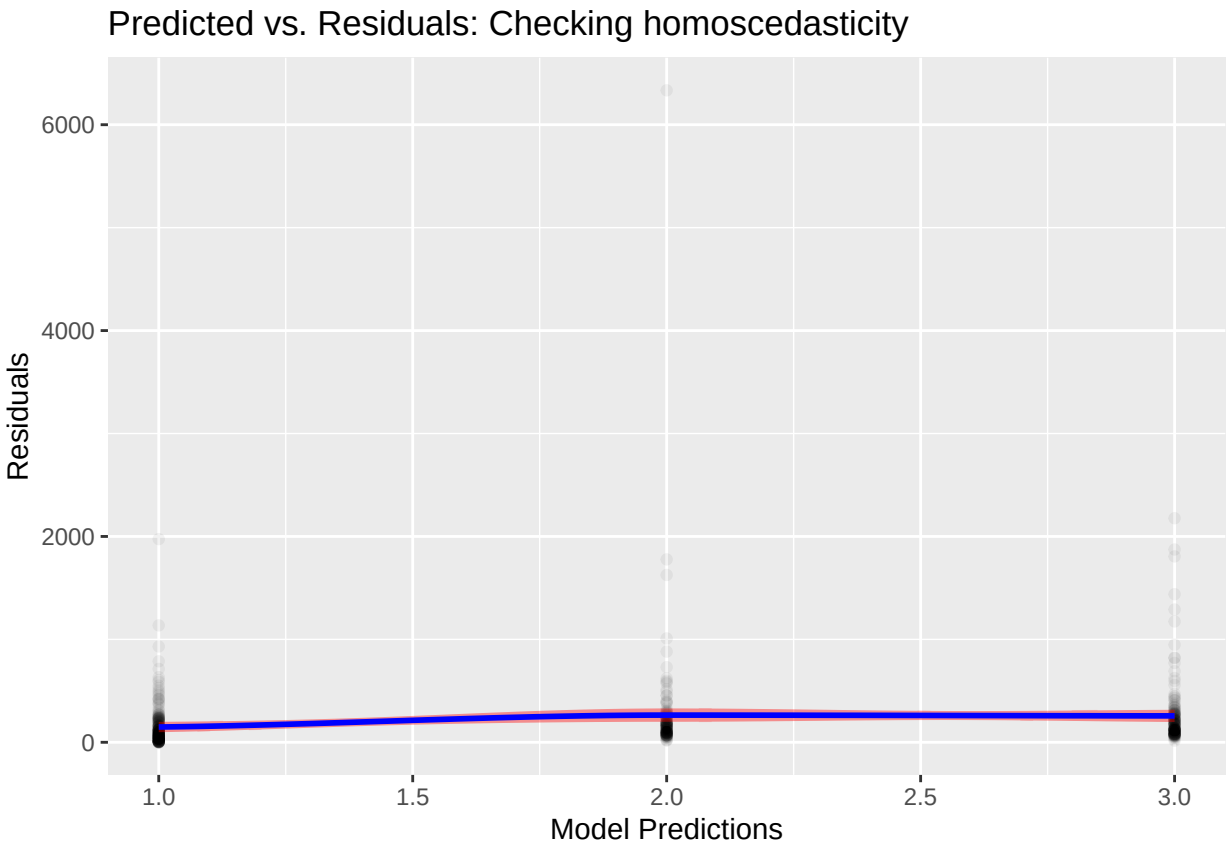
5a. Large-Sample Assumptions As a team, evaluate all of the large sample model assumptions. However, you do not necessarily want to discuss every assumption in your report. Instead, highlight any assumption that might pose significant problems for your analysis. For any violations that you identify, describe the statistical consequences. If you are able to identify any strategies to mitigate the consequences, explain these strategies.

The sample size of 500 observations indicates that we must evaluate and consider the following large-sample assumptions as pertaining to our model.

1. Data is Independent and Identically Distributed (I.I.D)

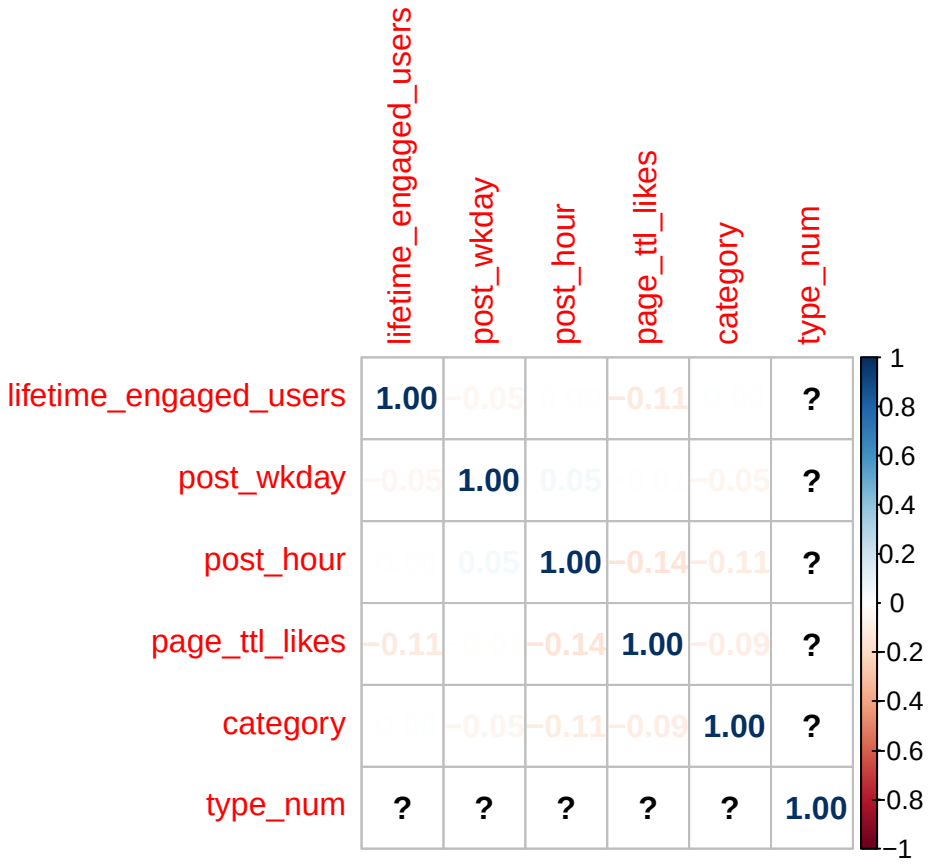
It is evident that our posts dataset is not I.I.D. As all of the posts are being made by the same cosmetics company and on the same platform (Facebook), inevitably all posts will be linked to the same company's products and promotions. Although we do not have unique detail regarding each post, it's likely that multiple posts may be tied to the same product or promotion. Other post parameters (time/date) may also not be independent, depending on the marketing strategy the team employed at the time this study was performed. Lastly, the number of impressions on each post are not independent from one another as many of the engagements/interactions may be from the same followers of the page across multiple posts.

Unfortunately, there is not much we can do to improve the raw data itself to reduce I.I.D. concerns. As all of the posts are anonymous we can't directly discern how to cluster or aggregate datapoints. Using robust standard errors may mitigate non-I.I.D. concerns and account for any heteroscedasticity in our model, although this is not an ideal solution.



2. Unique BLP Exists - No perfect collinearity

To satisfy this assumption, no variable in our model can be written as a linear combination of the other variables in the model. This assumption is satisfied given the following correlation matrix (modifying the "type" covariate to numerical):



Intentionally omitted variables

As with all real-world datasets, there are many variables that were intentionally or unintentionally omitted from our final model. Some variables were accessible within the dataset but were omitted, including:

- Lifetime post total reach
- Lifetime post total impressions
- Lifetime post impressions by people who have liked the page
- Lifetime post reach by people who have liked the page
- Lifetime people who have liked page and engaged with post
- Comments
- Likes
- Shares

As our outcome variable *engaged users* is defined as the number of users who clicked on the post in question, we can see that many of the intentionally omitted variables are very similar in what they measure - most are defined as some metric of engagement. As we determined that *engaged users* would be the best-suited outcome variable for our model, as it was a more holistic summation of engagement than individual metrics such as *comments* or *post reach only by people who liked the page*.

These variables that could be classified as outcome variables were omitted, as including these in the model would obscure the effects and relationships of the true measured variables. For example, we could have included "likes" as a covariate and examined its effect on user engagement. However, "likes" and engagement are directly related, and including this parameter in our model would absorb causal effect from other variables that we are trying to measure.

Unintentionally omitted variables

Many other variables that may affect user engagement were not tracked in this dataset, as a result they are

not able to be factored into our model. We would like to address some of these omitted variables below and discuss their potential impact on the model results.

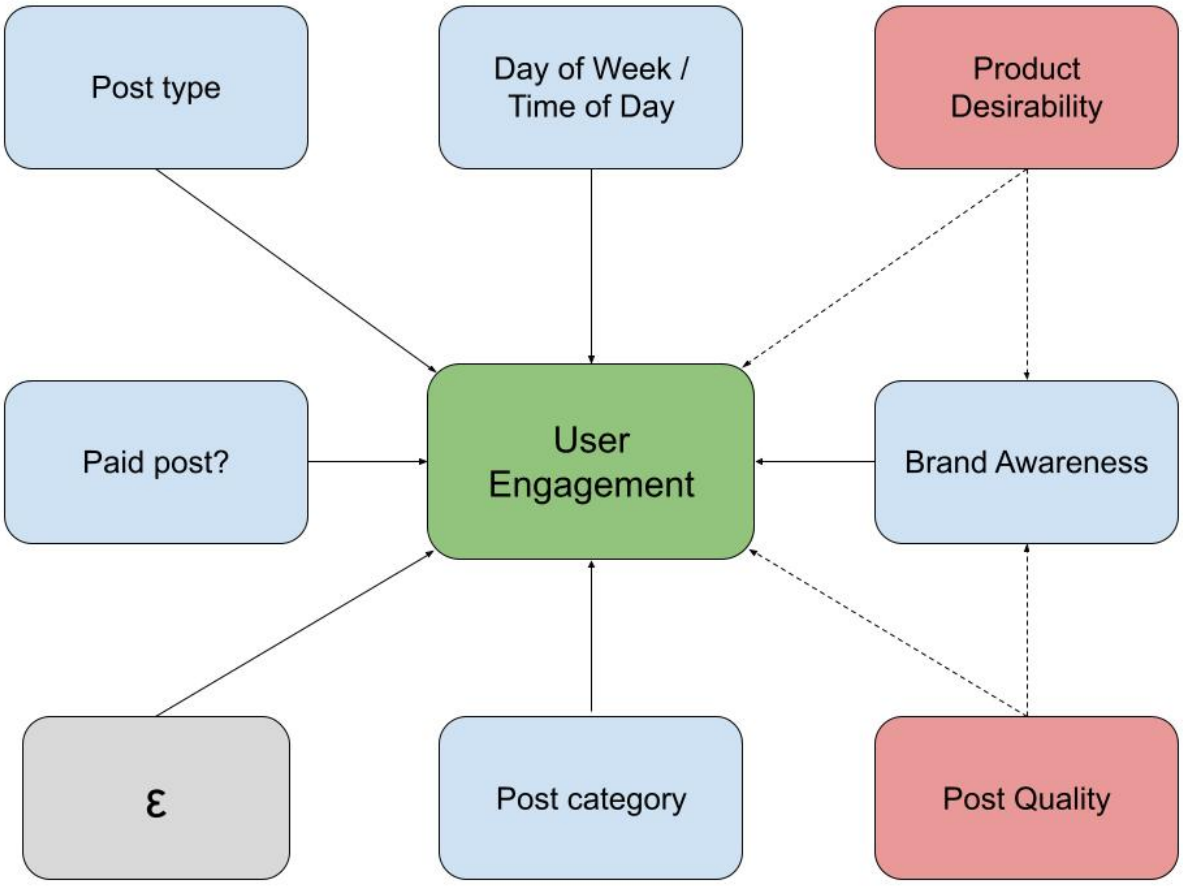
- Desirability of the product being marketed:

Not all products that the company promotes on its page will be of equal desirability to its customers. For example, two posts of the same type and category (not to mention all of the other supporting covariates) could have wildly varying engagement figures depending on if the product was highly in-demand or not. If we had knowledge of the product/service being promoted in each post, we would be able to account for it in our model by factoring in parameters such as the sales figures of each product, market sentiment of each product, etc. (this information could likely be obtained from external sources). We would then be able to add sales figures or another product desirability metric into our model; promoting a more well-known product would likely lead to more engagement than a more obscure product, all other factors constant.

- Quality of the promotional material:

Similarly, not all posts made by the company on the Facebook page will be of similar quality. Even if posts are classified here as "Photo" or "Video", posting a well-designed infographic may lead to more engagement than sharing an image of lower design quality. This may be hard to objectively quantify but if we had access to the contents of each post it's possible that we would be able to classify posts as "high quality", "low quality" etc. Incorporating this into the model may help the company decide how much to invest in its graphic design/marketing agencies - if there is a significant effect when a higher-quality post is used, we may be able to quantify the positive effect a higher-quality post has on user engagement and ultimately product revenue.

Both of these omitted variables ultimately affect the brand awareness of the company, as more desirable products and higher quality content both would increase traffic to the site. Currently our metric for brand awareness is the amount of likes on the page, but if these omitted variables were included in our analysis, potentially we could engineer a more holistic brand awareness metric as an input into our model. Both of these variables also directly have an effect on user engagement - the full causal graph of our model is shown below.



#

In summary, one of the main limitations of this dataset is the amount of information regarding the posts themselves - although we know the type and category of post, we don't have visibility to the nature of the post contents, which we assume can have an effect on the company's brand awareness as well as user engagement.

7. Conclusion

Make sure that you end your report with a discussion that distills key insights from your estimates and addresses your research question.

Encouragement for the Project

This project touches on many of the skills that you have developed in the course.

- When you are reasoning about the world and the way that it works, you are implicitly reasoning about *random variables*. Although you might not reason with specific functions (e.g. $f_x(x) = x^2$) to describe these random variables, you are very likely to be reasoning about conditional expectations.
- This class is not a class in pure theory! And so, theories you have about the world need to be informed by samples of data. These samples might be iid, or they might not be. The team will have to assess how this, and other possible violations of model assumptions shape what they learn.
- Given a set of input variables, OLS regression produces an estimate of the BLP. But, how good of a predictor is this predictor? And, does the team have enough data to rely on large-sample theory, or does the team need to engage with the requirements of the smaller-sample?
- Throughout, you will have to communicate both to a technical and non-technical audience.