Estimating Missing Data's Effects on Causal Inference with Diff-in-Diff and IP-weighting

Nathen Byford

2024-12-15

Introduction

Missing data is a common problem among statistical analyses. Data can be missing due to a variety of reasons, form a subject not answering a question, to a subject leaving a study for one reason or another. Sometimes missing data is numerous and other times a study can have no missing data. Often times when a study has plentiful missing values classical statistical methods using the complete cases will be biased and something is needed to be done.

In causal inference the issue of missing data is no different, there can be unintended bias introduced based on values that are missing. Causal inference methods might have more or less bias introduced by missing data due to the fact that we are trying to estimate counter factual outcomes, outcomes that don't exist in the first place. These estimates for the counter factual outcomes are based on the data observed in the sudy and if values are missing, information about the counter factuals is also being lost. Because of this I aim to investigate how causal inference estimates differ when there is missing data.

Missingness in Data

It is important to understand the different types of missing data that can emerge in studies. Missingness can be grouped into 3 main categories; Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) Little and Rubin (2019). The best case for missing data is that it is MCAR meaning that the data are missing with no pattern or undying process. Because MCAR data are assumed to be missing at random, using complete case analysis with classical statistical methods do not induce bias. Missing at random behind MCAR in the fact that it may be caused by something, but complete case or simple data imputation will produce similar results. MNAR is the worst case when there is some underlying process causing the missingness and a pattern to the missing data. In this case a stronger imputation method is required to reduce bias compared to complete case analysis which is not recommended for MCAR data.

Methods

This study looks into the differences in the estimated treatment effect for compete case analysis in MNAR data compared to the imputed data estimated treatment effects. The following subsections go into detail about the methods of data imputation and causal inference to estimate the treatment effect.

Data Imputation

Using Machine Learning techniques Haliduola, Bretz, and Mansmann (2022) are able to impute MNAR data from a clinical trial for anxiety medication. The first step of the imputation was to cluster the data by tox response. This response curve is used to group better understand the differences between subjects based on their initial and continued response to the drug. Due to the time component of the data a recursive neural network was utilized in the data imputation. In addition due to the small sample size of some cluster over sampling was used in the training dataset. Because of this method, the data that are MNAR can be imputed with minimal loss of information and induced bias.

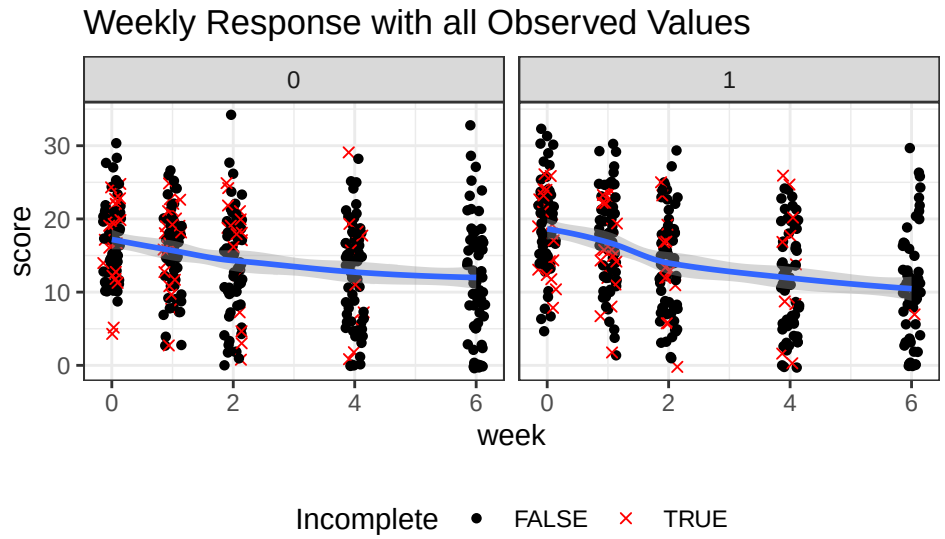
Causal Inference Methods

Two causal inference methods will be used to estimate the treatment effect of the anxiety drug from the baseline checkup to the final checkup. The two methods that will be tested are difference in difference (Diff-in-Diff) and IP-weighting. Both methods estimate the average treatment effect over time assusing that the assumptions hold.

Analysis

Exploratory Data Analysis

The response variable is based on the Hamilton Anxiety Rating Scale (HAMA), therefore a lower score represents a better response. These scores where observed at a baseline at week 0 and then after treatment at week 1, 2, 4, and 6. Below in figure 1 we can see the scores of each subject at each checkup, untreated is on the left and treated subjects are on the right. The observations shown as a red "X" are ones that contain missing score values.



Treatment	NA count
1	38
0	42

Week	NA count
0	0
1	0
2	14
4	23
6	43

Complete Case analysis

Using the complete cases in the dataset 80 observations are lost due to missing response values.

Imputed Values analysis

Conclusion and Discussion

References

Haliduola, Halimu N., Frank Bretz, and Ulrich Mansmann. 2022. "Missing Data Imputation in Clinical Trials Using Recurrent Neural Network Facilitated by Clustering and Oversampling." *Biometrical Journal* 64 (5): 863–82. https://doi.org/10.1002/bimj.202000393.

Little, Roderick J. A., and Donald B. Rubin. 2019. *Statistical Analysis with Missing Data*. John Wiley & Sons.