Estimating how Missing Data Affects Causal Inference with Diff-in-Diff and IP-weighting

Nathen Byford

2024-12-12

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. Missingess can be grouped into 3 main categories; Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR).

Methods

Analysis Results

Conclusion and Discussion