IMS Health & Quintiles are now IQVIA™

Propensity scores and missing data

28 March 2019

Quratul Ann

Table of Contents

Introduction

- The aim of propensity score (PS) is to estimate the effects of a treatment in observational studies we need to be able to make fair comparisons between treated and untreated individuals

- Treated and untreated individuals are likely to be different in many ways

- We need to measure these differences and account for them in some way in the analysis

- This method was first introduced in 1983 by Rosenbaum and Rubin.

- Since, the propensity score methods have been increasingly used to adjust for confounding in many fields

Assumptions

- To estimate a causal effect using propensity score methods, we typically make four key assumptions

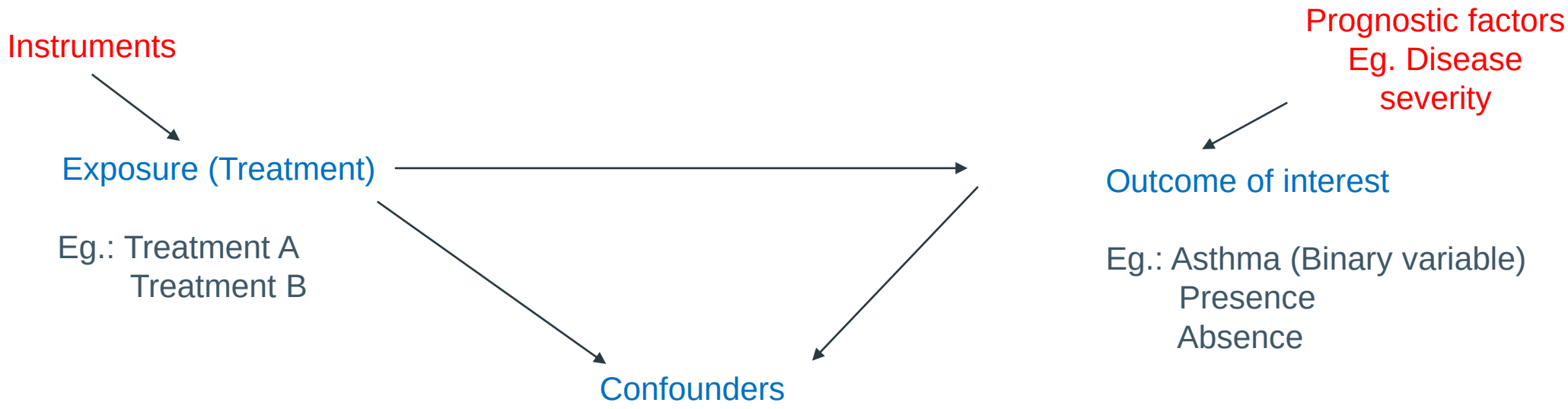
    1. Positivity
    2. Consistency
    3. No interference
    4. No unmeasured confounding

Note:

Whenever we are trying to attach “causal” interpretations to any sort of regression model, we are typically making these assumptions, although they are rarely stated explicitly.

Factors included (Confounders) in propensity score

How to choose which variables to include in the propensity score model?



Variables associated with:

- Exposure and outcome
- Outcome only
- Exposure only

Consider for propensity score?

- Yes
- Yes – may improve efficiency
- Maybe not – possibility to increase the variance (accounting for the uncertainty)

Estimating Propensity scores

- Propensity scores are predicted probabilities
- Calculate a propensity score for receiving a treatment for each patient, whether they take a treatment or not based on risk factors for receiving the drug
- Uses logistic regression with treatment (exposure status) as the outcome and covariates (possible confounders) as the exposures

Propensity score methods

- Propensity score matching
  - Estimand to estimate (ATT or ATE)
  - 1:1 Matching, 1:4 Matching, many to one and so on
  - With replacement or without replacement
  - Nearest Neighbour – with a calliper of 0.01
  - Kernel and Local linear

- Propensity score stratification
- Propensity score covariate adjustment
- Inverse probability of treatment weighting using the propensity score

How do propensity scores compare with multivariable analyses?

- Results are generally very similar
- Propensity scores +
  - Useful where large numbers of confounders present
  - Useful when rare outcome (standard regression models may not converge)
  - Allow identification of comparable groups
- Multivariable analysis +
  - Shows separate effect of each covariate
- Both methods –
  - Still don’t account for unknown confounding!

Missing data in propensity score analysis

- The problem
  - Ps analysis might be biased if missing data are ignored

- Potential solutions:
  - Complete case analysis
  - The missingness pattern approach (MPA)/missing category
  - Multiple imputation

Complete Case Analysis (CCA)

- PS is estimated only for patients for whom **all the variables are observed: even the outcome** (even though Y not used at this stage)
- This is because we want to achieve the covariates balance between groups of subjects included in the analysis
- Outcome model estimated **on the same sub-sample** of patients
- CCA is biased if the **distribution of the confounders is distorted** among complete cases
- However, as long as missingness **does not depend both on the outcome variable and treatment variable**, this bias is generally small if:
  - variables associated with missingness are not strong confounders
  - missingness rate is low

The missing pattern approach (MPA)

- Method proposed by d'Agostino and Rubin
- Definition of a **Generalized PS** estimated **within each pattern** of missingness
- Relies on the **validity of a set of assumptions** that cannot be expressed using Rubin's taxonomy (MCAR/MAR/MNAR)
- No **unmeasured confounding** when considering each missingness pattern separately
- The PS estimated using the MPA **will balance the observed part of the confounders only**
- Need the **treatment variable to be fully observed**: often the case in Electronic Health Records (EHR) data
- The number of different PS models needs to be equal to the number of different patterns of missingness
- MPA becomes challenging when the **number of patterns of missingness increases**: pooling patterns?
- **Bootstrap** for the confidence intervals

The missing indicator method

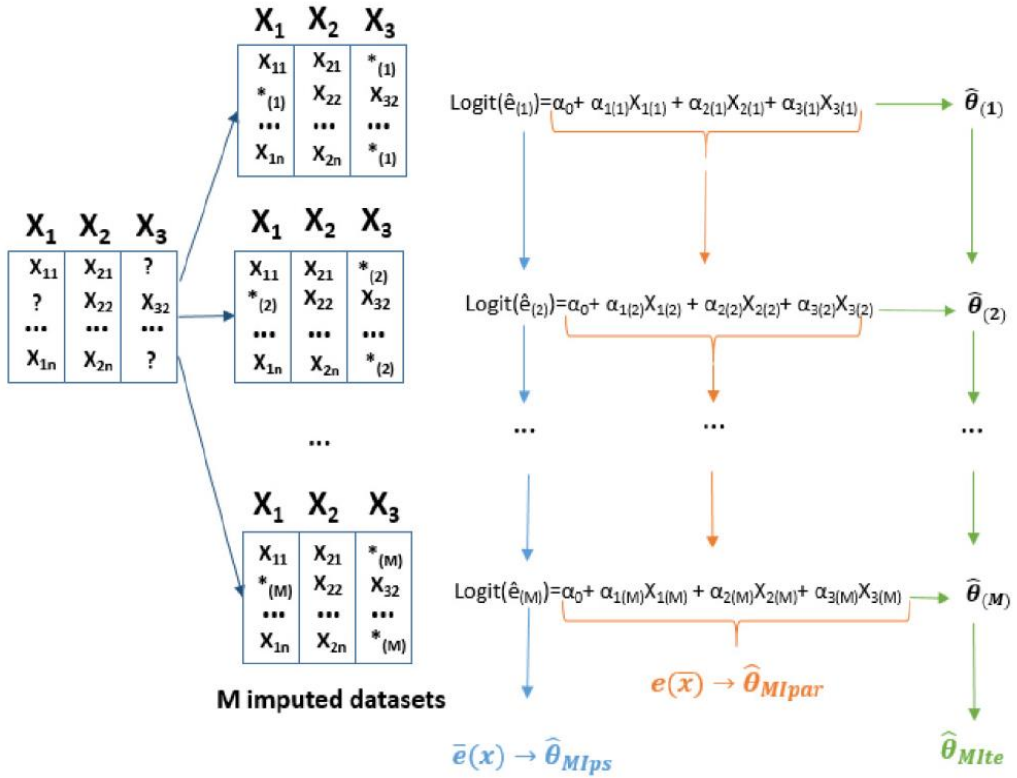
- **For categorical variables:** a “missing” category is added to the variable when used in the PS model
- **For continuous variables:** a constant value replaces missing data and a missing indicator is added in the PS model
- Straightforward and very common approach:

  ⟶ **when is it valid?**

- Link to the MPA
- When there is only 1 confounder:
  - Both methods are equivalent. Therefore, the missing indicator method **relies on the same assumptions**
- When there are additional (fully observed) confounders:
  - The missing indicator approach is a **simplification of the MPA**
  - Which assumes no interaction between the missingness indicators and the fully observed covariates
- The MPA relies on fewer assumptions

Multiple imputation

- **Aim**: create *M* complete datasets to estimate the PS for each participant and apply Rubin's rules to obtain a treatment effect estimate



- Only **MIte is a consistent** estimator of the treatment effect*
- The full analysis (PS model+outcome model) must be performed **within each dataset**
- Then, the *M* estimated treatment effects are **pooled using Rubin's rules**

Shall we consider the outcome variable in the imputation model?

YES, otherwise the true association between outcome and exposure variable is not reflected among imputed values and this will lead to **biased estimates**

Missing at random (MAR) assumption need to be satisfied

Summary of key assumptions

- **Complete Case Analysis (CCA)**
  - Missingness doesn't depend on outcome variable (or on both outcome and treatment variable if estimating OR), and
  - Variables associated with missingness are not strong confounders
- **Missingness Pattern Approach (MPA)**
  - Missing values arise only in confounders
  - Partially missing variables are only confounders when observed
- **Missingness indicator method**
  - Same as MPA but also…
  - The effect of fully observed confounders on propensity to receive treatment is the same, irrespective of which partially-observed confounders are missing
- **Multiple imputation (MI)**
  - Missing At Random (MAR)

Key advantages and disadvantages

- **Complete Case Analysis (CCA)**
  - Pro: Simple, easy
  - Con: Lower precision
- **Missingness Pattern Approach (MPA)**
  - Pro: Retains full sample in the analysis
  - Con: Need treatment and outcome to be fully observed
  - Con: Difficult if there are many different missingness patterns
- **Missingness indicator method**
  - Pro: Simple. Removes difficulty of many missingness patterns
  - Con: Need treatment and outcome to be fully observed
- **Multiple imputation (MI)**
  - Pro: Powerful and can handle missing values in confounders, treatment, and outcome
  - Con: Doesn’t mix well with PS matching
  - Con: Implementation more complex than other methods