Missing data in propensity score analysis

Clémence Leyrat, Elizabeth Williamson

Department of Medical Statistics, LSHTM
Outline

• The problem

• Missing data mechanisms

• Potential solutions:
  • Complete case analysis
  • The missingness pattern approach/missing category
  • Multiple imputation

• Summary
Why do we care?

• As for any other statistical method, PS analysis might be biased if missing data on confounders are ignored.

• In PS analysis:

<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$\hat{e}(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$x_{11}$</td>
<td>$x_{21}$</td>
<td>$x_{31}$</td>
<td>$\hat{e}_1(x)$</td>
</tr>
<tr>
<td>0</td>
<td>$x_{12}$</td>
<td>?</td>
<td>$x_{32}$</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>$x_{13}$</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>0</td>
<td>$x_{14}$</td>
<td>$x_{24}$</td>
<td>$x_{34}$</td>
<td>$\hat{e}_4(x)$</td>
</tr>
<tr>
<td>1</td>
<td>$x_{15}$</td>
<td>$x_{25}$</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

PS often used when the number of potential confounders is quite large and low proportion of complete cases.
In practice

Despite STROBE recommendations, half of the studies did not report how missing data were handled.

Missingness mechanisms

• The impact of missing data on the treatment effect estimate depends on the reason for missingness

• Rubin’s taxonomy:
  • **MCAR**: the probability of data being missing does not depend on the values observed or unobserved variables
  • **MAR**: the probability of data being missing does not depend on the unobserved data, conditional on the observed data
  • **MNAR**: the probability of data being missing depends on the unobserved data, even after conditioning on the observed data
Choosing a missing data method

• The choice of the appropriate method to handle missing confounder data in PS analysis depends on the **missingness mechanisms**

• DAGs can be useful to assess whether the **estimand of interest** can be recovered from the observed data:

Complete case analysis (CCA)
CCA in PS analysis

- PS 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 balance the covariates between groups of subjects included in the analysis.

- Outcome model estimated on the same sub-sample of patients
CCA in regression analyses

• CCA unbiased under MCAR and some MAR scenarios in regression models

• CCA might be valid in regression if missingness does not depend simultaneously on the outcome Y and the treatment Z

• This is true when estimating a conditional effect but PS are used to estimate marginal effects
Implications

• CCA biased if the distribution of the confounders is distorted among complete cases

• However, as long as missingness does not depend both on Y and Z, this bias is generally small

• OK if:
  • variables associated with missingness are not strong confounders
  • missingness rate is low
The missingness pattern approach (MPA)
Principle

• Method proposed by d’Agostino and Rubin

• Definition of a generalized PS estimated within each pattern of missingness

<table>
<thead>
<tr>
<th>Z</th>
<th>x_1</th>
<th>x_2</th>
<th>x_3</th>
<th>( \hat{\epsilon}(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>x_{11}</td>
<td>x_{21}</td>
<td>x_{31}</td>
<td>( \hat{\epsilon}_1(x) )</td>
</tr>
<tr>
<td>0</td>
<td>x_{12}</td>
<td>?</td>
<td>x_{32}</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>x_{13}</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>0</td>
<td>x_{14}</td>
<td>x_{24}</td>
<td>x_{34}</td>
<td>( \hat{\epsilon}_4(x) )</td>
</tr>
<tr>
<td>1</td>
<td>x_{15}</td>
<td>x_{25}</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X_2</th>
<th>X_3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>( \hat{\epsilon}(X_1, X_2, X_3) )</td>
<td>( \hat{\epsilon}(X_1, X_2) )</td>
</tr>
<tr>
<td>Missing</td>
<td>( \hat{\epsilon}(X_1, X_3) )</td>
<td>( \hat{\epsilon}(X_1) )</td>
</tr>
</tbody>
</table>

Perils and Promises of Propensity Scores – 2nd June 2020
Assumptions

• Validity of MPA relies on the validity of a set of assumptions that cannot be expressed using Rubin’s taxonomy (MCAR/MAR/MNAR)

• Let $X$ the vector of baseline confounders be split in $X = (X_{\text{obs}}, X_{\text{mis}})$ and $R$ the vector of the missingness indicators for the confounders.

• We need first the validity of a SITA-type assumption
  $$(Y(0), Y(1)) \perp Z \mid X, R$$
Assumptions (2)

And either:
\[ X_{\text{mis}} \perp Z \mid X_{\text{obs}}, R \quad \text{or} \quad X_{\text{mis}} \perp (Y(0), Y(1)) \mid X_{\text{obs}}, R \]

What does that mean?

X can be a confounder when observed but not when missing

Examples

• Example: CKD stage and AKI

• Example: Quality of life

Why? The PS estimated using the MPA will balance the observed part of the confounders only
Implementation

• Need the **treatment variable to be fully observed**: often the case in EHR data

• The number of different PS models is 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
Example (1)

- **Data source**: UK Clinical Practice Research Datalink. n=570,586.

- **Exposure**: new prescription of antihypertensive drugs (ACEI/ARB)

- **Outcome**: acute kidney injury (AKI) within 5 years

- **Confounders**: gender, age, ethnicity, prescription of other antihypertensive drugs, chronic comorbidities at start of follow-up, including chronic kidney disease (CKD)

- **Missing data**:  
  - CKD stage: 52.9%  
  - Ethnicity: 59%  
  - Complete cases: n=121,527
Example (2)

- **CCA**: huge loss of sample size (efficiency) and risk of bias

- **Multiple imputation (MI)**: valid under a MAR mechanism. Here: baseline CKD stage is more likely to be recorded for patients with a lower level of kidney function → MNAR

- **MPA?** If CKD stage not available, unobserved information cannot be used to determine the GP’s treatment decision to prescribe ACEI/ARBs.
Example (3)

Results (IPTW):

<table>
<thead>
<tr>
<th>Method</th>
<th>Risk difference (per 1000)</th>
<th>Normal-based bootstrap 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crude effect (no adjustment)</td>
<td>13.30</td>
<td>12.52-14.08</td>
</tr>
<tr>
<td>Complete case analysis</td>
<td>4.60</td>
<td>2.76-6.45</td>
</tr>
<tr>
<td>Missingness pattern approach</td>
<td>5.96</td>
<td>5.10-6.82</td>
</tr>
</tbody>
</table>
The missing indicator method
Principle

• **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
### Example

#### Results (IPTW):

<table>
<thead>
<tr>
<th>Method</th>
<th>Risk difference (per 1000)</th>
<th>Normal-based bootstrap 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crude effect (no adjustment)</td>
<td>13.30</td>
<td>12.52-14.08</td>
</tr>
<tr>
<td>Complete case analysis</td>
<td>4.60</td>
<td>2.76-6.45</td>
</tr>
<tr>
<td>Missingness pattern approach</td>
<td>5.96</td>
<td>5.10-6.82</td>
</tr>
<tr>
<td>Missing indicator approach</td>
<td>5.93</td>
<td>5.01-6.85</td>
</tr>
</tbody>
</table>
Multiple imputation
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.

- Two key questions when using MI in the PS context:
  
  Should the outcome be included in the imputation model?

  =>$\text{PS paradigm} \neq \text{Missing data paradigm}$

  How to apply Rubin's rules?

  =>$\text{pooled treatment effect or pooled PS?}$
What should be pooled?

• Only **multiple imputation (MI)** is a consistent estimator of the treatment effect*

• As expected, 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**:
  • Overall treatment effect = average of the $M$ values
  • Variance: pooled variance accounting for the uncertainty in the imputation of missing values

* C. Leyrat et al. Propensity score analysis with partially observed covariates: How should multiple imputation be used? Statistical Methods in Medical Research. 2017 DOI: 10.1177/0962280217713032
Issue with matching

• In PS matching, **some patients are discarded** from the analysis if no good match cannot be found

• When using MI, matched patients will be different across imputed datasets => **relevance of the pooled estimate**?

• But remember that pooling the PS to match only once is not appropriate
Specification of the imputation model

• Should we **include the outcome** in the imputation model?  
  **YES** otherwise the true association between Y and X is not reflected among imputed values

• Chained equations or joint modelling?  
  **MI using chained equations** is more flexible (different model for each partially observed variable)
Specification of the imputation model

• Should we include interactions or quadratic terms?
  If such effects exist, yes, but the specification of the imputation model can be challenging to ensure compatibility with the analysis model
  => Substantive model compatible imputation (smcfcs)*

• How many imputed datasets?
  Depends on the proportion of missing data, but at least 10

Example

**Study**: Effect of statins on short term mortality after pneumonia

**Population**: n=7158 patients with a diagnosis of pneumonia of whom 1398 were under statin treatment without CHD

**Outcome**: Mortality, $P_0 = 0.37$, $P_1 = 0.22$

**Covariates**: 22 confounders, 3 of them partially observed:
- BMI missing for 19.2% of patients
- Smoking status missing for 6.2%
- Alcohol consumption missing for 18.5%
- Complete cases: n=5168
PS distribution (on CC)
Example - results

Risk ratios

<table>
<thead>
<tr>
<th>Method</th>
<th>Initial data</th>
<th>Increasing the missingness rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RR</td>
<td>95% CI</td>
</tr>
<tr>
<td>Crude</td>
<td>0.590</td>
<td>[0.531;0.652]</td>
</tr>
<tr>
<td>CC</td>
<td>0.703</td>
<td>[0.582;0.850]</td>
</tr>
<tr>
<td>MP</td>
<td>0.719</td>
<td>[0.609;0.850]</td>
</tr>
<tr>
<td>MIte</td>
<td>0.666</td>
<td>[0.662;0.670]</td>
</tr>
<tr>
<td>MIps</td>
<td>0.665</td>
<td>[0.661;0.669]</td>
</tr>
<tr>
<td>MIpar</td>
<td>0.658</td>
<td>[0.654;0.662]</td>
</tr>
</tbody>
</table>

MI: 10 imputed datasets using chained equations. No interactions or quadratic terms in the imputation model

Data assumed to be MAR

3 MI: small differences because the 3 variables were not very strong confounders

MPA: some patterns were pooled because of too few participants
Conclusion

• The choice of the method to handle missing data in PS analysis depends on the reasons for missingness

• No “universal method”

• When publishing the results of a PS analysis:
  • Give the % of missing data for each variable
  • Discuss the possible reasons for missingness to justify the choice of the method to handle it
  • Describe the method clearly (e.g., for MI, specification of the imputation model, number of imputations...)

Perils and Promises of Propensity Scores -- 2nd June 2020