Graphical Criteria of Recoverability under Not Missing at Random Case

Authors: Jiwoo Kim[1*], Felix Thoemmes[1]

[1]Department of Psychology, Cornell University

E-mail: jk2759@cornell.edu Webpage: https://jiwookim-ci.github.io/

Abstract

Missing data is a well-known source of bias in research, with Not Missing at Random (NMAR) cases being particularly challenging due to the dependence between missingness itself and target variables with missing values. This dependence makes it generally impossible to recover target estimands without bias. This study aims to provide a generalizable framework for determining the recoverability of statistical parameters under NMAR conditions. Specifically, we introduce criteria that classify NMAR models based on the recoverability of key parameters, such as the slope and intercept in linear models. To achieve this, we use causal graphical models and path tracing rules to establish systematic criteria for identifying different types of bias. Building on previous work by Mohan and Pearl (2021) and Thoemmes and Mohan (2015), we utilize a new type of Missingness-Directed Acyclic Graph (m-DAG). By applying path tracing rules in this m-DAG, we derive conditions under which parameters of interest in linear models can be recovered without bias. This study contributes to the methodological literature by offering a structured approach for researchers to assess the impact of NMAR on their analyses and determine whether unbiased estimates can still be obtained under NMAR conditions.

Notation

Independent variable: X. Dependent variable: Y

Missing indicators for X and Y: R_x, R_Y

Corresponding observed variables, whose values are deterministically set to be either identical to X or Y, respectively, or set to missing: X_{obs}, Y_{obs}

Introduction

Background & Motivation

- Under Not Missing at Random (NMAR), missingness depends on unobserved values or unmeasured variables and modern methods like multiple imputation or MLE cannot eliminate bias under NMAR.
- Previous studies focused on how NMAR data are generated, not on bias or recoverability and few general rules exist for when and how parameters can be recovered under NMAR [1,2].

Aims of this study

- Propose **graphical criteria** for analyzing NMAR using causal diagrams to check recoverability of intercepts and slopes in linear models [3].
- Help applied researchers visualize assumptions and determine when estimates are unbiased.

Missingness-Directed Acyclic Graph

Conditioning vs Controlling

Conditioning:

- Analyzes only a stratified subset of the data.
- Blocks confounding paths by subsetting a confounder.
- In missing data, this aligns with observing only where data are not missing.
- Notation in DAG: double circle

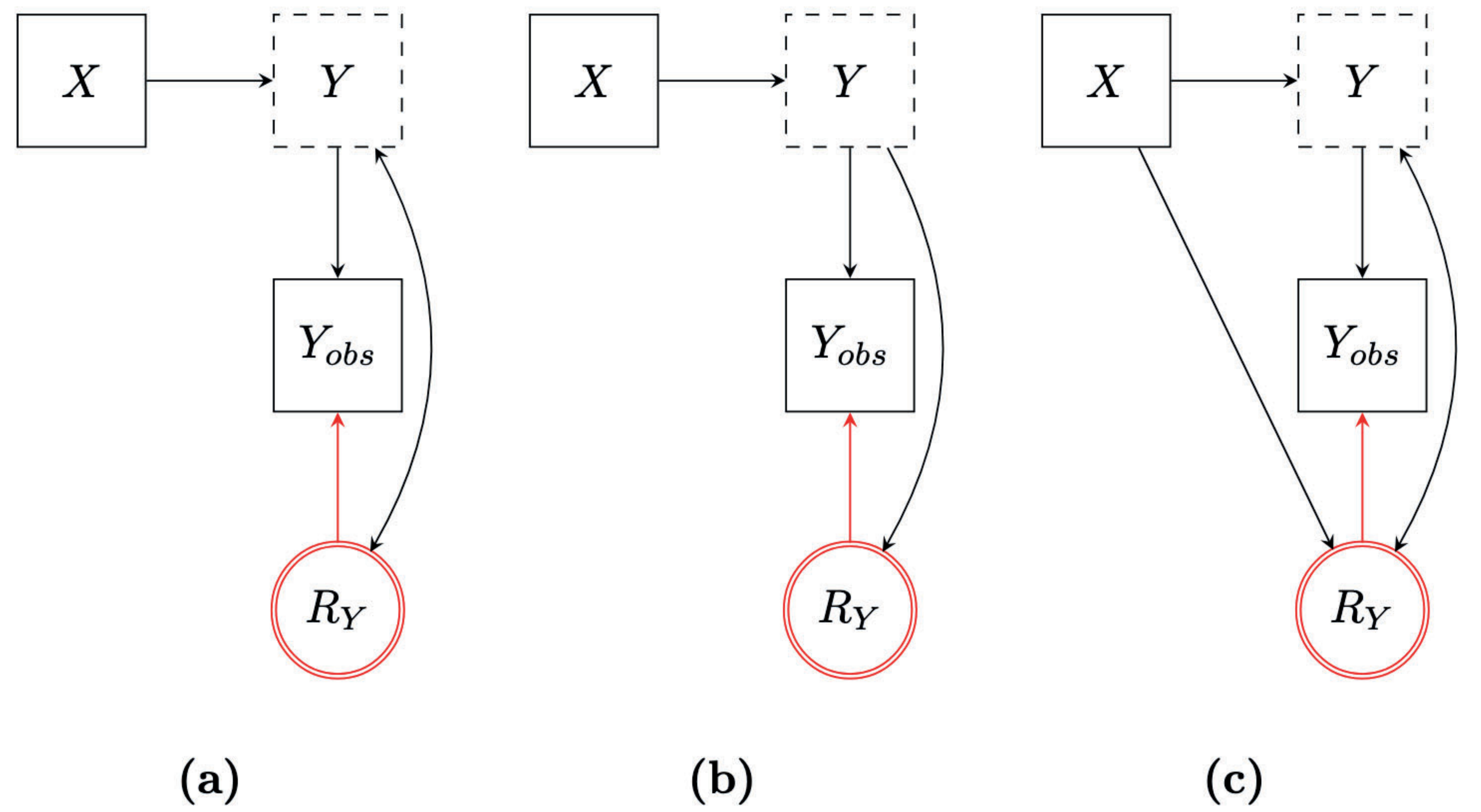
Controlling:

- Uses the entire dataset, adjusting for confounders via regression or similar models.
- Adds predictors to control for confounding rather than subsetting.
- Notation in DAG: square

→ **Missingness is a case of conditioning**, given that only parts of the X or Y data can be used for analysis.

In Figure 1, R_Y in (a), (b), and (c) are conditioned. The red path from R_Y to Y_{obs} generated by conditioning on the R_Y variable, as Y_{obs} is a deterministic variable of Y and R_Y.

Figure 1



Graphical criteria

A researcher first needs to define the target estimand and the target variables that are involved in the definition of this target [4].

- Focal variables (intercepts) : Y, Y_{obs}
- Focal variables (slopes) : X, Y

(1) bias in intercepts

- Bias occurs if any path from Y to Y_{obs} passes through the missingness indicator R_Y. This applies to:
- Model (a) – Open paths: $Y \rightarrow Y_{obs}$, $Y \leftrightarrow R_Y \rightarrow Y_{obs}$
- Model (b) – Open paths: $Y \rightarrow Y_{obs}$, $Y \rightarrow R_Y \rightarrow Y_{obs}$
- Model (c) – Open paths: $Y \rightarrow Y_{obs}$, $Y \leftrightarrow R_Y \rightarrow Y_{obs}$, $Y \leftarrow X \rightarrow R_Y$

In all three models, means are biased.

(2) bias in slopes

- Bias in the slope between X and Y occurs if conditioning on R_Y opens a backdoor or collider path. This applies to:
- Model (a) – Conditioning on R_Y does not bias slope (R_Y is neither a collider nor a descendant of Y)
- Model (b) – Conditioning on R_Y biases slope (R_Y is a descendent of Y)
- Model (c) – Conditioning on R_Y biases slope (R_Y is a collider between X and Y)

In model (b) and (c), slopes are biased

Graphical criteria - bias in intercepts

For each missingness indicator $R \in \{R_X, R_Y\}$

1.Check whether $Y \perp\!\!\!\perp R \mid X$

- If this d-separation statement is true, there is no bias.
- If this d-separation statement is false and the variables are d-connected, the intercept of the parametric model will be biased

Graphical criteria - bias in slopes

For each missingness indicator $R \in \{R_X, R_Y\}$

1. Check whether $Y \perp\!\!\!\perp R \mid X$

- If this d-separation statement is true, there is no bias in the slope of the parametric model.
- If this d-separation statement is false, proceed to the next step.

2. Check whether $X \not\!\perp\!\!\!\perp R \mid Y$ and $X \perp\!\!\!\perp R$:

- If <u>**both**</u> of these d-separation statements are true, there is bias only in the ***intercept***.
- If at least one of the two d-separation statements is false, bias occurs in both ***intercept*** and ***slope***.

Conclusion

- We propose general criteria for identifying which parameters are recoverable without bias under NMAR.
- Using m-DAGs with distinct notations for conditioning and controlling, we clarify mechanisms behind bias in NMAR models.
- Our approach allows researchers to assess recoverability of means and slopes (e.g., treatment effects), even under listwise deletion.

References

1. Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592. https://doi.org/10.2307/2335739
2. Thoemmes, F., & Mohan, K. (2015). Graphical representation of missing data problems. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(4), 631-642. https://doi.org/10.1080/10705511.2014.937378
3. Mohan, K., & Pearl, J. (2021). Graphical models for processing missing data. *Journal of the American Statistical Association, 116*(534), 1023-1037. https://doi.org/10.1080/01621459.2021.1874961
4. Quimpo, N. J., & Steiner, P. M. (2024, October 18). Drawing Directed Acyclic Graphs for Causal Inference. https://doi.org/10.31234/osf.io/u4yta