Lab 11: Missing Data

PHS2000B, Spring 2023

Rienna Russo

April 18, 2023

Learning Objectives

1. **Key Concepts and Terminology in Missing Data**
2. **Structural Representations of Selection Bias**
3. **Structural Representations of Missing Data**
4. **Methods to Address Missing Data**
5. **More DAGs :)**

Thanks Emma!

Huge thanks to Emma McGee whose slides served as the basis for today's lab!

Key Concepts and Terminology in Missing Data

Missing Data

Missing data are present in all areas of population health research:

- Respondents **do not answer** certain survey questions
- Individuals are **lost to follow-up** or **miss specific study visits**
- Lab samples are **unusable** or **below the limit of detection**
- Certain measures **cannot be routinely collected** for all study participants
- Data files are **inaccessible, corrupted, or lost**
- Etc.

Missing data may result in loss of **statistical efficiency** and potentially **biased estimates**.

Notation Review

We begin by defining a set of **notation.**

Notation	Definition
Z	A vector of all variables of interest in a given analysis, including the outcome, exposure(s), and all covariates
Z^{obs}	A (sub) vector of variables within Z which are **observed** for a given individual i
Z^{mis}	A (sub) vector of variables within Z which are **missing** (unobserved) for a given individual i
R	A missing indicator vector which has elements that take the value 1 for all Z^{obs} (observed covariates) and 0 for all Z^{mis} (missing covariates)

Example: We are interested in analyzing the vector of variables $Z = \{$ age, highest level of education, earnings $\}$. For an arbitrary individual i, let $Z_i = \{30, ., .\}$ where . indicates a missing value. Thus, for this individual i, $R = \{1, 0, 0\}$.

Classification of Missing Data

In their canonical work on missing data, Little and Rubin[1,2] classified missing data into three categories:

1. **Missing Completely At Random (MCAR)**
2. **Missing At Random (MAR)**
3. **Missing Not At Random (MNAR or NMAR)**

We can formally define MCAR, MAR, and MNAR using our new set of notation.

[1] Rubin DB (1976). Inference and missing data. Biometrika 63:581-592.
[2] Little RJA, Rubin DB. (1987) Statistical Analysis with Missing Data.New York: John Wiley Sons.

Classification of Missing Data: MCAR

Data are defined as **Missing Completely At Random (MCAR)** if the probability of the data being observed is independent of both observed and unobserved data:

$$Pr(\boldsymbol{R}|\boldsymbol{Z}) = Pr(\boldsymbol{R}|\boldsymbol{Z}^{\boldsymbol{obs}}, \boldsymbol{Z}^{\boldsymbol{mis}}) = Pr(\boldsymbol{R}|\boldsymbol{\theta})$$

Where $\boldsymbol{\theta}$ represents the set of parameters that govern the probability distribution of the missing indicator vector, R.

In other words, data are MCAR when the **missingness is entirely due to random chance.**

Classification of Missing Data: MAR

Data are defined as **Missing At Random (MAR)** if, conditional on the observed data, the probability of the data being observed is independent of the unobserved values of Z:

$$Pr(\boldsymbol{R}|\boldsymbol{Z}) = Pr(\boldsymbol{R}|\boldsymbol{Z}^{obs}, \boldsymbol{Z}^{mis}) = Pr(\boldsymbol{R}|\boldsymbol{Z}^{obs}, \boldsymbol{\theta})$$

That is, data are MAR if, **conditional on observed values of $\boldsymbol{Z}$, missingness is random.**

MCAR is a special case of MAR because under MCAR:

$$\boldsymbol{R} \perp\!\!\!\perp \boldsymbol{Z}^{obs} \implies Pr(\boldsymbol{R}|\boldsymbol{Z}^{obs}, \boldsymbol{\theta}) = Pr(\boldsymbol{R}|\boldsymbol{\theta})$$

Classification of Missing Data: MNAR or NMAR

Data are defined as **Missing Not At Random (MNAR)** if the probability of the data being observed is a function of values of Z that *would have been observed*:

$$Pr(\boldsymbol{R}|\boldsymbol{Z}) = Pr(\boldsymbol{R}|\boldsymbol{Z}^{\boldsymbol{obs}}, \boldsymbol{Z}^{\boldsymbol{mis}}, \theta)$$

That is, **conditional on the observed values of Z, missingness depends on unobserved values of Z.**

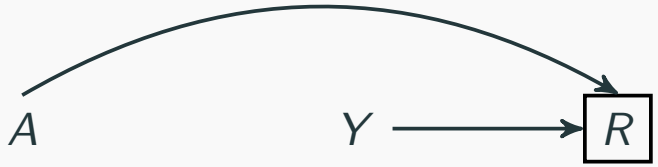
We cannot use our empirical data alone to determine whether the data are MAR or MNAR.[3]

[3] An empirical distinction between MNAR vs. MAR would require access to data which *were not observed.*

Structural Representations of Selection Bias

Structural Approaches to Selection Bias

More recent literature has focused on **structural representations of missing data** (and also of **selection bias** more broadly).[4] For example, some of the earliest causal DAGs you saw probably depicted a simple selection bias scenario resulting from **collider stratification:**



Where A is the exposure, Y is the outcome, and R is an indicator for selection into the study.

[4] Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004 Sep;15(5):615-25

Revisiting Selection Bias

In some settings, bias due to conditioning on having observed (non-missing) data may be represented by collider stratification on a causal DAG. However, bias due to missing data may also be present even in the absence of collider stratification! To understand why this is the case, we will start by providing a very **general definition of selection bias:**

> **Selection bias:** The parameter of interest in the target population differs from the parameter in the subset of individuals from the population available for analysis.[5]

[5] Many epidemiologists use the term selection bias to refer exclusively to biases that affect internal validity. The definition proposed here encompasses biases which can result from either a lack of internal or external validity. This approach is aligned with a recent body of literature on *target validity* (see Supplemental Slides).

A Note on Selection Bias vs. Selection

Structural representations can also help us distinguish between selection vs. selection bias:

Selection:

The observed study sample is not representative of the target population of interest. Whether or not bias is induced depends on the estimand we wish to estimate.

Selection Bias

The target estimand of interest is not equal to the estimate obtained from the observed data as a result of selection. Results in biased estimates.

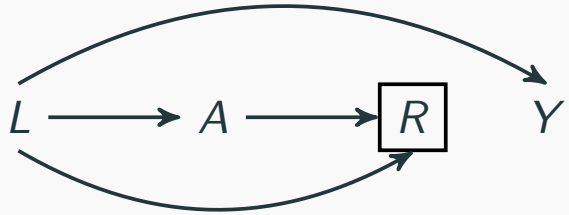
* this encompasses collider stratification, which is the usual structural selection bias and also failure to generalize

Structural Representations of Selection Bias with Collider Stratification

We can now differentiate **two potential sources of selection bias.**

1. **Selection bias due to collider stratification:** Conditioning on a collider (or a descendent of a collider) on a path between exposure A and outcome Y generally induces a non-causal $A - Y$ association, even if the exposure has no causal effect on the outcome (i.e., *selection bias under the null*).[6]

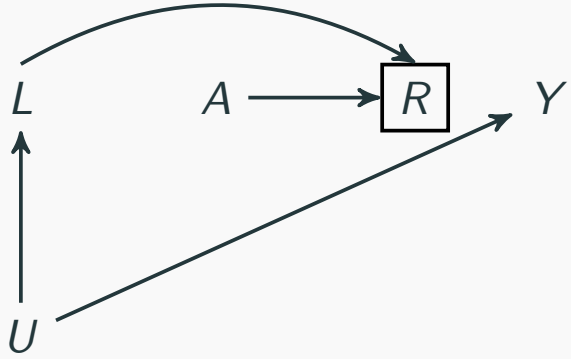
For example, consider this DAG where R represents missing data and L is a confounder:



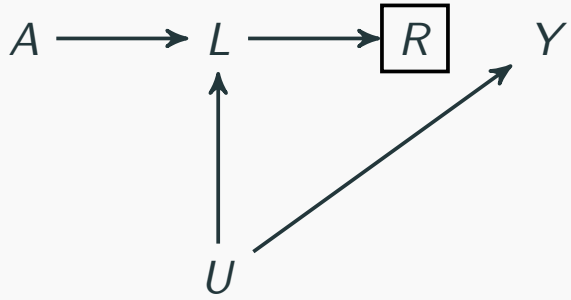
[6] Although collider stratification always induces an association between the causes of the collider, that association may in some cases be restricted to specific levels of the collider.

Structural Representations of Selection Bias with Collider Stratification

a) Selection bias is induced through the $A \leftarrow R \rightarrow L \leftarrow U \rightarrow Y$ path since we are conditioning on R, which is a collider.



b) Selection bias is induced through the $A \leftarrow L \leftarrow U \rightarrow Y$ path since we are conditioning on R, which is a descendant of a collider, L.



Structural Representations of Selection Bias without Collider Stratification

In the absence of collider stratification, selection bias is *not guaranteed* to arise. However, under certain conditions, selection bias *may occur in the absence of colliders.*

2. **Selection bias without colliders:** Selection occurs when the observed study sample is not equal to the target population of interest. When selection is present but there are no colliders, the estimated quantity may be unbiased for the subset of the population that was selected but *biased in the entire study population*. This form of selection bias essentially occurs when the selected sample is not representative of the target population $\implies$ **failure to generalize.** This concept applies to both descriptive and causal measures.[7,8]

We will now review *two separate scenarios* that can lead to selection bias without colliders.

---

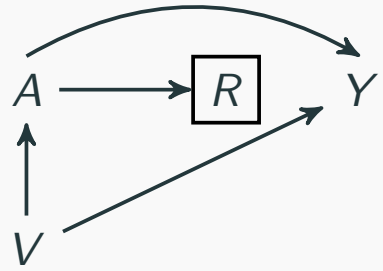
[7] Greenland S. Response and follow-up bias in cohort studies. Am J Epidemiol. 1977 Sep;106(3):184-7.

[8] Hernán MA. Invited Commentary: Selection Bias Without Colliders. Am J Epidemiol. 2017 Jun 1;185(11):1048-1050.

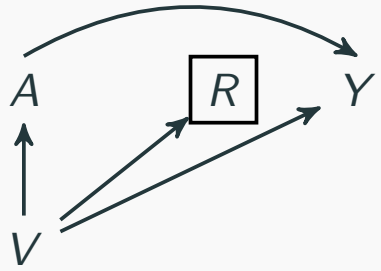
Selection Bias without Collider Stratification: Causal Effects

i. **Selection bias for a causal effect without colliders:** Selection bias may arise in the absence of colliders when, on the scale of interest for the analysis, the *effect of the exposure on the outcome is heterogeneous across some level(s) of V*, where V represents a third variable that is associated with selection. Sometimes, such bias only arises when the *exposure has a non-null effect on the outcome.* For example:[9]

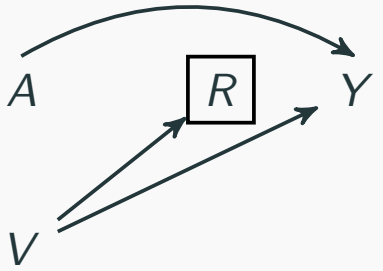
a)



b)



c)



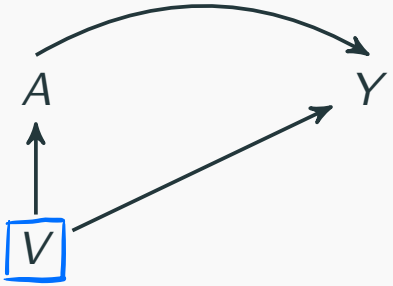
[9] These DAGs were adapted from Smith, L.H. Selection Mechanisms and Their Consequences: Understanding and Addressing Selection Bias. Curr Epidemiol Rep 7, 179–189 (2020).

Selection Bias without Collider Stratification: Causal Effects

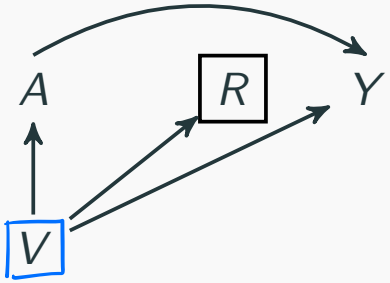
CONDITIONAL EFFECT: $E[Y|A=1,V=v] - E[Y|A=0,V=v] = E[Y|A=1,V=v,R=1] - E[Y|A=0,V=v,R=1]$

b/c $Y \perp\!\!\!\perp R \mid V$ since conditioning on V blocks the path from Y to R so we can remove R from the conditioning statement based on the independence property, e.g. $Pr(A|B,C) = Pr(A|B)$ if $A \perp\!\!\!\perp C \mid B$

DAG A: No Selection



DAG A: Selection



NOTE: If we have EMM, our conditional effect may be unbiased but our marginal estimate in our study sample to be a biased estimate for the marginal effect in the total population
⇒ FAILURE TO GENERALIZE!

MARGINAL EFFECT: $\sum_v \left(E[Y|A=1,V=v] - E[Y|A=0,V=v]\right) Pr(V=v) \neq \sum_v \left(E[Y|A=1,V=v,R=1] - E[Y|A=0,V=v,R=1]\right) Pr(V=v|R=1)$

For this to be true we need BOTH $Y \perp\!\!\!\perp R|V$ and $V \perp\!\!\!\perp R$.

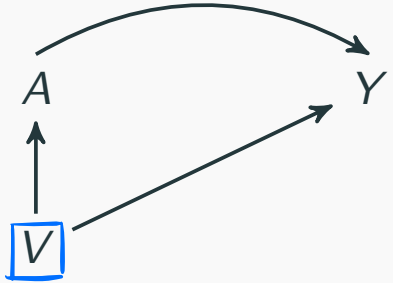
$V \perp\!\!\!\perp R$ when there is no open path from V to R. This is when $Pr(V=v) = Pr(V=v|R=1) = Pr(V=v|R=0)$

Since there is an arrow from V to R, $V \not\perp\!\!\!\perp R$ and these quantities are not equivalent

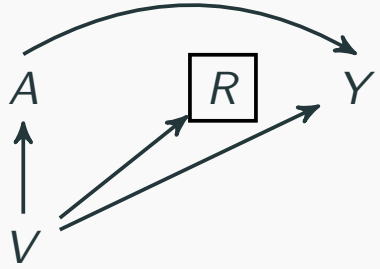
The exception is if there is no EMM. When we don't have EMM then our conditional effect is equal to the marginal effect and it doesn't matter that $V \not\perp\!\!\!\perp R$ since the distribution of V doesn't influence the effect

Selection Bias without Collider Stratification: Causal Effects

DAG A: No Selection



DAG A: Selection



Even though $E[Y|A = a, V = v] = E[Y|A = a, V = v, R = 1]$ for both exposure values, we expect the distribution of V to differ in the selected population, since V influences R: $V \not\perp\!\!\!\perp R$ and $Pr[V = v] \neq Pr[V = v|R = 1]$.

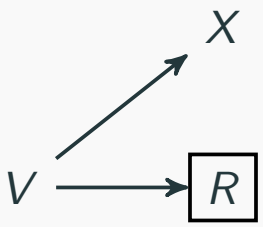
Unless the causal effect on the scale of interest is homogeneous across values of V, the average effect in the population will differ from the effect in the selected population.

see previous slide for more details!

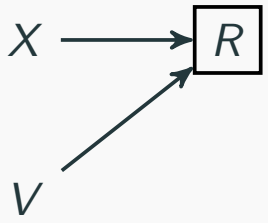
Selection Bias without Collider Stratification: Descriptive Measures

ii. **Selection bias for a descriptive measure:** Say we are interested in describing the distribution of some variable X. Non-random selection into the study may prevent a descriptive measure calculated in the selected population from being *generalizable* or *transportable* to the target population. For example, each of the following DAGs represent a scenario in which selection could bias our descriptive estimate of X:[10]

a)



b)

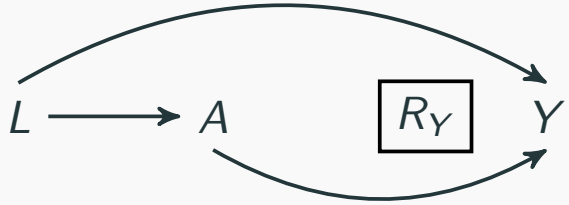


[10] For an excellent overview on the use graphical models such as DAGs in descriptive surveys, see: Schuessler J, Selb P. Graphical causal models for survey inference [Internet]. 2019.

Structural Representations of Missing Data

Structural Representations of Missing Data: MCAR Example

These same structural conclusions apply when selection is specifically due to **missing data.** For example, consider the following DAG where we are interested in estimating the causal effect of A on Y, $\boldsymbol{Z}^{\boldsymbol{obs}} = \{L, A\}$, and $\boldsymbol{Z}^{\boldsymbol{mis}} = \{Y\}$ (i.e., we are missing data on the outcome Y):

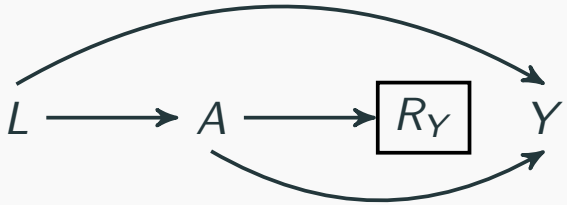


Data on Y are MCAR because $R_Y \perp\!\!\!\perp (Y, A, L) \implies R_Y \perp\!\!\!\perp (\boldsymbol{Z}^{\boldsymbol{mis}}, \boldsymbol{Z}^{\boldsymbol{obs}})$.[11]

[11] We saw in lecture that it can be helpful to subscript the missing indicator R by the variable(s) which are missing (e.g., R_Y if Y is the only variable with missing values). We will use this notation in the following DAGs.

Structural Representations of Missing Data: MAR Example

Now let's consider an alternative DAG where we assume that missingness in the outcome Y depends on the exposure A. Once again $\boldsymbol{Z}^{obs} = \{L, A\}$ and $\boldsymbol{Z}^{mis} = \{Y\}$:



Data on Y are MAR because $R_Y \perp\!\!\!\perp Y|(A, L) \implies R_Y \perp\!\!\!\perp \boldsymbol{Z}^{mis}|\boldsymbol{Z}^{obs}$. If the effect of A on Y varies across levels of L, then this is an example of selection bias without colliders! * FOR MARGINAL EFFECT

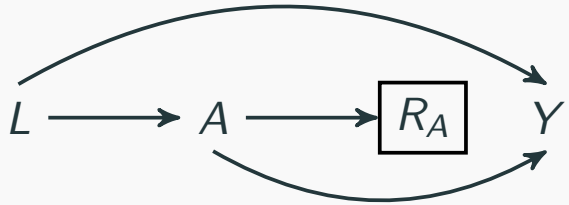
CONDITIONAL: $E[Y|A=1, L=\ell, R_Y=1] - E[Y|A=1, L=\ell, R_Y=1]$

MARGINAL: $\sum_{\ell} E[Y|A=1, L=\ell, R_Y=1] - E[Y|A=1, L=\ell, R_Y=1] \Pr(L=\ell|R_Y=1)$

since $L \not\perp\!\!\!\perp R_Y$ then $\Pr(L=\ell|R_Y=1) \neq \Pr(L=\ell)$ and if we have EMM our **MARGINAL** effect will be biased ("FAILURE TO GENERALIZE")

Structural Representations of Missing Data: MNAR Example

Finally, let's consider a DAG where we assume that there is missing data for the variable A and missingness depends on A itself. Now we have $\boldsymbol{Z}^{\boldsymbol{obs}} = \{Y, L\}$ and $\boldsymbol{Z}^{\boldsymbol{mis}} = \{A\}$:



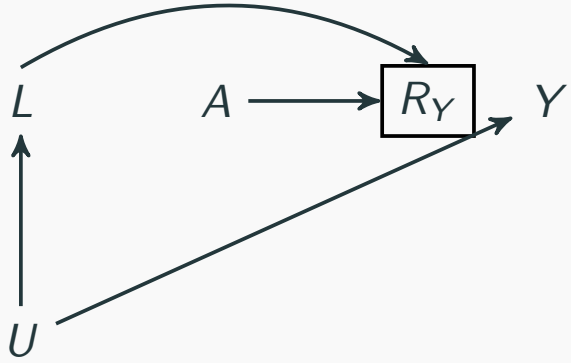
Data on A are MNAR because $R_A \not\perp\!\!\!\perp A|(Y, L) \implies R_A \not\perp\!\!\!\perp \boldsymbol{Z}^{\boldsymbol{mis}}|\boldsymbol{Z}^{\boldsymbol{obs}}$. We will return to this MNAR example after reviewing methods which can be used to address missing data.

Structural Representations of Missing Data

NOTE: DAGs are drawn under the null (i.e., no arrow from A to Y)

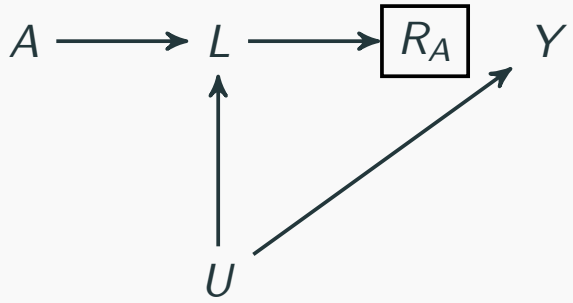
Question: What kinds of missingness are represented by these DAGs? Check $R \perp\!\!\!\perp Z^{mis} | Z^{obs}$

a)



$R_Y \perp\!\!\!\perp Y | A, L$? Yes! Data are MAR

b)



$R_A \perp\!\!\!\perp A | L, Y$? Yes!

Colliders are path specific so conditioning on L (and R_A) opens $A \rightarrow \boxed{L} \leftarrow U \rightarrow Y$ path but blocks the path $A \rightarrow \boxed{L} \rightarrow R_A$

Methods to Address Missing Data

Methods to Address Missing Data

Several methods can be used to **address missing data,** including:

- **Ad Hoc Procedures** (e.g., last observation carried forward (LCOF))
- **Complete Case Analysis**
- **Multiple Imputation**
- **Weighting and Model-Based Approaches**

Whether a given method (e.g., complete case analysis) leads to bias will depend on whether the **proposed estimator** provides an **unbiased estimate** of the **estimand of interest** *under a given set of assumptions.*

Ad Hoc Procedures

Procedure	Description	Major Disadvantages
Available cases	For each analysis, include all cases where the variable(s) involved in that analysis are present	Sample base may change across different analyses
Drop covariates	Completely drop covariate(s) with large amounts of missing data	May result in uncontrolled confounding
Missing indicator	Create a missing value indicator which is included in the analytic model(s)	Makes very strong assumptions; biased even under MCAR
Single imputation	"Fill in" each missing value with a single imputed value, then analyze the imputed dataset using standard methods	Ignores uncertainty in the imputed missing values; may be biased even under MCAR (e.g., LOCF)

Complete Case Analysis

In a **complete case analysis,** only those observations for which *all variables are measured* are analyzed. This method is simple but **inefficient.** Complete case analyses **may also be biased:**

Data Are:	Does Complete Case Analysis Induce Bias?
MCAR	No
MAR	It depends. In general, if missingness depends on exposure A and/or measured covariates L but not on the outcome $Y \implies$ unbiased estimates. If missingness depends on the outcome $Y \implies$ biased estimates. In time-varying settings, complete case analyses will also often result in biased estimates.
MNAR	It depends. If missingness depends on a variable U which was not measured at all $\implies$ biased estimates. In some cases, complete case analyses may provide biased estimates for the marginal causal effect of interest in the target population but unbiased estimates for the conditional causal effect in the selected population. In the absence of effect modification on the scale of interest, an unbiased conditional causal effect would be equal to the marginal causal effect.

Complete Case Analysis

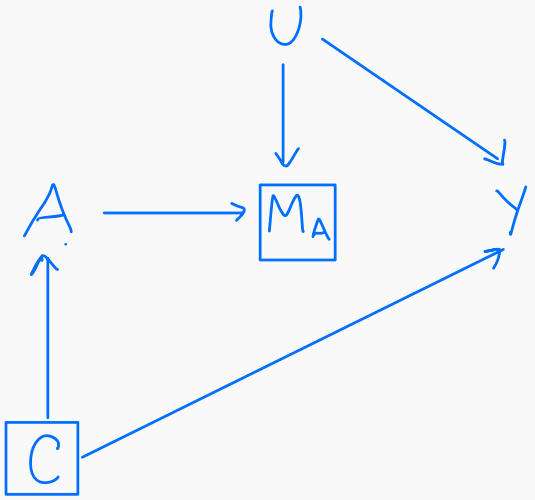
The potential for bias in a complete case analysis also **depends on the outcome model.** For example, complete case analyses are more robust to bias under logistic outcome regression:[12]

	Variable(s) Missingness Depends On:	Linear Regression	Logistic Regression
1	None (i.e., data are MCAR)	Unbiased	Unbiased
2	Outcome	Biased[a]	Unbiased
3	Exposure (and possibly confounders)	Unbiased	Unbiased
4	Outcome and confounders	Biased	Unbiased
5	Outcome and exposure (and possibly confounders)	Biased	Biased[b]

[12] Adapted from Hughes RA, Heron J, Sterne JAC, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. Int J Epidemiol. 2019 Aug 1;48(4):1294-1304. [a] Biased in general, except when in truth there is no association between the outcome and the exposure. [b] Biased in general, except when missingness depends on the outcome and exposure independently.

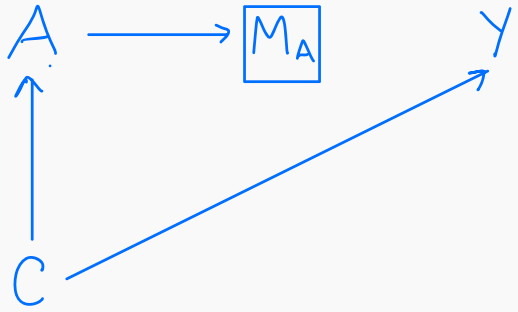
Breaking Down The Table (Next Slide)

example from class with MNAR



Complete case analysis will be biased

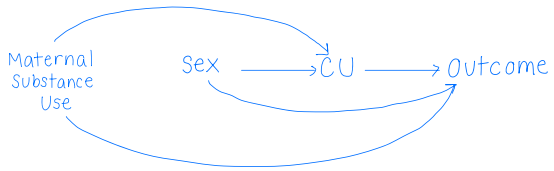
* open path from R_A to Y through U



Complete case will be unbiased since there are no open paths

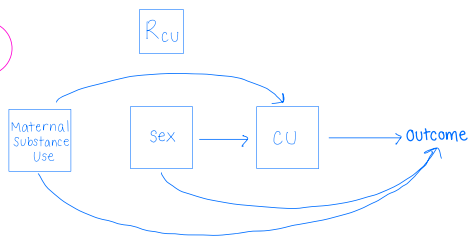
FROM HUGHES 2019 PAPER

ORIGINAL DAG



Given missingness in the exposure, we expect the exposure coefficient to have bias based on the DAG and type of model used. CU = cannabis use, exposure of interest; R_{CU} = missingness in CU

1



Missing Completely At Random

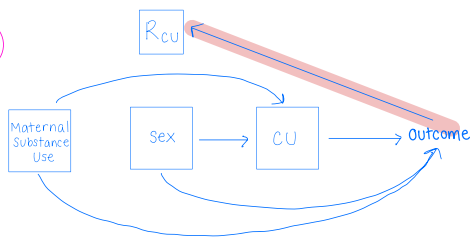
Linear: Unbiased

Logistic: Unbiased

No paths from R_{CU} to any variable

* $CU \perp\!\!\!\perp R_{CU} \Rightarrow$ MCAR → CATE is UNBIASED

2



MAR: R_{CU} depends on outcome

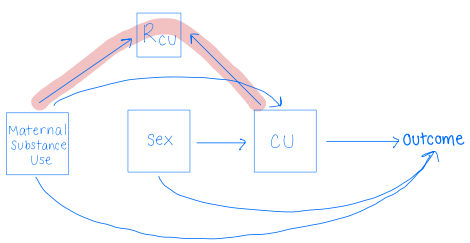
Linear: Biased

Logistic: Unbiased b/c symmetry

Direct path from outcome to missingness

* $R_{CU} \perp\!\!\!\perp CU \mid$ (Outcome, Sex, Maternal Substance Use) ⟶ MAR
* Outcome $\not\perp\!\!\!\perp R_{CU} \mid$ CU, Sex, Maternal Substance Use) ⟶ CATE is BIASED

3



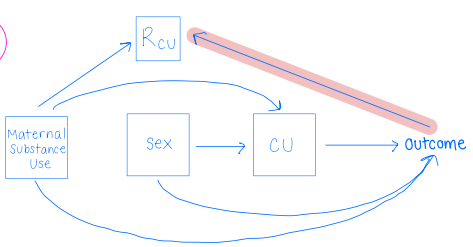
MNAR: R_{CU} depends on exposure

Linear: Unbiased

Logistic: Unbiased

* $R_{CU} \not\perp\!\!\!\perp CU \mid$ (Outcome, Sex, Maternal Substance Use) ⟶ MNAR
* Outcome $\perp\!\!\!\perp R_{CU} \mid$ CU, Sex, Maternal Substance Use) ⟶ CATE is UNBIASED
  ↳ Special case of MNAR when conditioning on covariates yields unbiased estimate since back door paths are blocked

4



MAR: R_{CU} depends on outcome

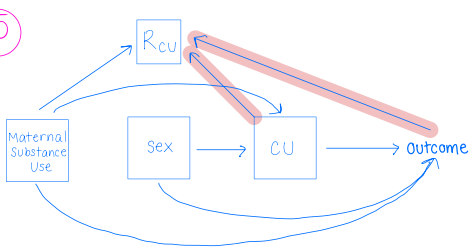
Linear: Biased

Logistic: Unbiased

Direct path from outcome to missingness

* $R_{CU} \perp\!\!\!\perp CU \mid$ (Outcome, Sex, Maternal Substance Use) ⟶ MAR
* Outcome $\not\perp\!\!\!\perp R_{CU} \mid$ CU, Sex, Maternal Substance Use) ⟶ CATE is BIASED

5



MNAR: R_{CU} depends on exposure and outcome

Linear: Biased

Logistic: Biased

Open path from exposure to outcome through R_{CU}

* Usual collider stratification selection bias → CATE is BIASED

Multiple Imputation

Multiple imputation essentially "fills in" (i.e., imputes) missing data using a pre-defined procedure **multiple times.**

- Multiple imputation is unbiased **under the assumption that data are MAR (or MCAR)** and there is no (gross) model mis-specification, where relevant.
- This approach has several key strengths, including:
  - An **increase in statistical efficiency** (as compared to a complete case analysis)
  - Explicit **incorporation of uncertainty in the imputation process**
  - **Ability to incorporate auxiliary variables** not included in the primary analysis

Multiple Imputation

Several different **imputation procedures** can be used to obtain multiply imputed datasets. Two general approaches are:

1. Imputing missing data by assuming that the missing and non-missing data follow a specific joint distribution **(Joint Modeling Procedure)** $\Longrightarrow$ **Amelia II**
2. Imputing missing data on a variable-by-variable basis by specifying a series of conditional densities **(Fully Conditional Specification)** $\Longrightarrow$ **Multiple Imputation with Chained Equations (MICE)**

Details on MICE and Amelia II are provided in the Supplemental Slides.

Weighting and Model-Based Procedures

Finally, let's review two additional approaches which can be used to address missing data.

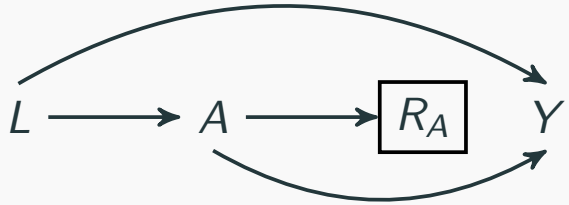
1. **Weighting procedures:** involve weighted complete case analyses where the weights are defined as the inverse of the probability of being observed (non-missing). This procedure essentially simulates what would have been observed if no one had been missing (non-stabilized weights) or if missingness had been random with respect to the covariates used in the missingness model (stabilized weights). Weighting is conducted under the assumption that the **data are MAR.** AKA IPW! Can think of this like censoring IPW
2. **Model-based procedures:** involve defining a model for the missing data and then conducting analyses and inferences based on that model. These approaches may provide unbiased results **even if the data are MNAR,** provided the missing data model is correctly specified. One example of such a procedure are **Heckman Selection Models**.

More DAGs :)

Structural Representations of Missing Data: MNAR Example

Let's use DAGs to think about how two of the methods we have just reviewed - **complete case analysis** and **multiple imputation** - would perform in the MNAR scenario we presented earlier.

Recall that we are interested in the causal effect of A on Y, there is missing data for the variable A, and missingness depends on A itself:



Data on A are MNAR because $R_A \not\perp\!\!\!\perp A|(Y, L) \implies R_A \not\perp\!\!\!\perp \boldsymbol{Z}^{\boldsymbol{mis}}|\boldsymbol{Z}^{\boldsymbol{obs}}$.

Structural Representations of Missing Data: MNAR Example

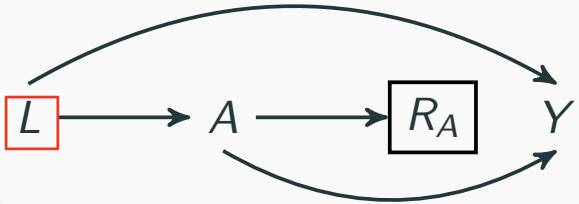
STEP 1: Determine missingness structure: $R_A \perp\!\!\!\perp A \mid (L, Y)$? No → MNAR

STEP 2: Address confounding bias → condition on L to block back door path $A \leftarrow L \rightarrow Y$

STEP 3: Assess for collider stratification bias (structural selection bias) → not a concern here

Question: Would a complete case analysis result in an unbiased estimate of the *conditional* effect of A on Y in the selected population? Under what conditions (if any) would it produce an unbiased estimate of the *marginal* effect of A on Y in the target population?

STEP 4: Assess for failure to generalize bias: check $Y \perp\!\!\!\perp R_A \mid L$ and $L \perp\!\!\!\perp R_A$



- $Y \perp\!\!\!\perp R_A \mid L$? Yes → Conditional effect unbiased

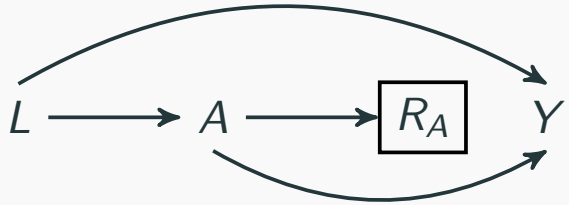
$$E[Y|A=1,L=\ell] - E[Y|A=0,L=\ell] = E[Y|A=1,L=\ell,R_A=1] - E[Y|A=0,L=\ell,R_A=1] \quad \text{b/c} \quad Y \perp\!\!\!\perp R_A \mid L$$

- $L \perp\!\!\!\perp R_A$? No
  - → With EMM: $\sum_\ell (E[Y|A=1,L=\ell] - E[Y|A=0,L=\ell]) \Pr(L=\ell) \neq \sum_\ell (E[Y|A=1,L=\ell,R_A=1] - E[Y|A=0,L=\ell,R_A=1]) \Pr(L=\ell \mid R_A=1)$ Marginal Effect Biased
  - → No EMM: $\underbrace{E[Y|A=1,L=0] - E[Y|A=0,L=0]}_{\text{CATE}(0)} = \underbrace{E[Y|A=1,L=1] - E[Y|A=0,L=1]}_{\text{CATE}(1)} = \underbrace{E[Y|A=1] - E[Y|A=0]}_{\text{ATE}}$ Marginal Effect Unbiased

Structural Representations of Missing Data: MNAR Example

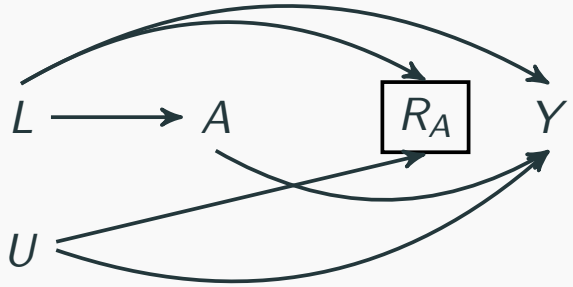
Question: Would multiple imputation result in an unbiased estimate of the *conditional* effect of A on Y in the selected population? Under what conditions (if any) would it produce an unbiased estimate of the *marginal* effect of A on Y in the target population?

We can't use multiple imputation since missingness depends on the variable missing



Structural Representations of Missing Data: Another MNAR Example

We also previously considered an MNAR example where missingness depended on a variable which was itself missing. Now, we assume that there is missing data for A and that missingness depends on L and on an unmeasured variable U. We have $\boldsymbol{Z}^{\boldsymbol{obs}} = \{Y, L\}$ and $\boldsymbol{Z}^{\boldsymbol{mis}} = \{A\}$:



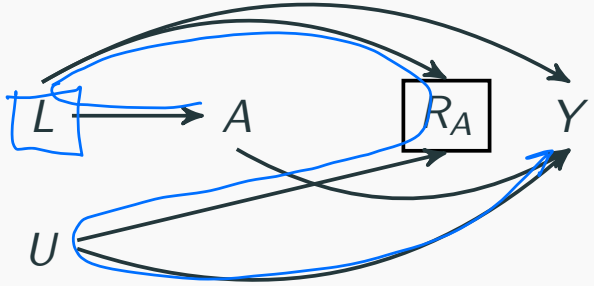
Data on A are MNAR because $R_A \not\perp\!\!\!\perp A|(Y, L) \implies R_A \not\perp\!\!\!\perp \boldsymbol{Z}^{\boldsymbol{mis}}|\boldsymbol{Z}^{\boldsymbol{obs}}$. Lack of independence arises as a result of collider stratification, where one of the causes of the collider, U, is unmeasured.

Structural Representations of Missing Data: Another MNAR Example

STEP 1: Determine missingness structure: $R_A \perp\!\!\!\perp A \mid (L, Y)$? No → MNAR

STEP 2: Address confounding bias → condition on L to block back door path $A \leftarrow L \rightarrow Y$

Question: Would a complete case analysis result in an unbiased estimate of the *conditional* effect of A on Y in the selected population? Under what conditions (if any) would it produce an unbiased estimate of the *marginal* effect of A on Y in the target population?



STEP 3: Assess for collider stratification bias (structural selection bias)

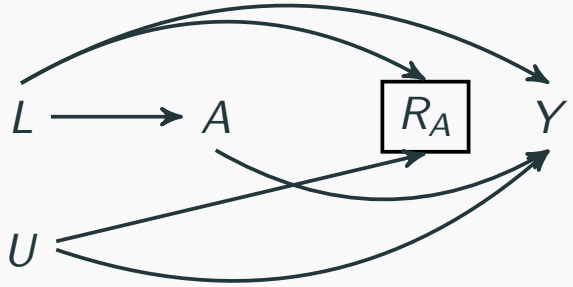
* We have an open path through R_A from A to Y! The unmeasured common cause, U, of R_A and Y is a collider, which opens path $A \leftarrow L \rightarrow [R_A] \leftarrow U \rightarrow Y$ BUT we can condition on L to block this path and the confounding $A \leftarrow L \rightarrow Y$ path

STEP 4: Assess for failure to generalize bias: check $Y \perp\!\!\!\perp R_A \mid L$ and $L \perp\!\!\!\perp R_A \rightarrow Y \not\perp\!\!\!\perp R_A \mid L$ biased conditional & marginal effects

Structural Representations of Missing Data: Another MNAR Example

Question: Would multiple imputation result in an unbiased estimate of the *conditional* effect of A on Y in the selected population? Under what conditions (if any) would it produce an unbiased estimate of the *marginal* effect of A on Y in the target population?

No → since missingness depends on an unobserved variable we can't use imputation



Supplemental Slides

Monotone Missingness

When multiple variables have missing values, the missingness is described as **monotone** if the variables (i.e., columns) can be rearranged such that observing one variable X_b for any individual i implies that X_a is observed for the same individual (where $a < b$).

	Monotone missingness				Non-monotone missingness			
Individual	Y	X1	X2	X3	Y	X1	X2	X3
1								
2								
3								
4								
5								
6								

Figure 1: Monotone missingness (left panel) vs. non-monotone missingness (right panel). Green cells denote observed data. Red cells denote missing data. Adapted from Horton and Kleinman (2007).

Unit vs. Item Non-response

Unit Non-Response: data on all covariates are missing for a given observation i at time t

- **Example:** a longitudinal household survey where household 2 did not provide a survey at time t

Survey responses at time t

ID	X1	X2	X3
1	40	1	2
2	.	.	.
3	47	0	3
4	52	0	2

Item Non-Response: one or more covariates (but not all) are missing for a given observation i at time t

- **Example:** a longitudinal household survey where X2 and X3 is missing for some households at time t

Survey responses at time t

ID	X1	X2	X3
1	40	.	2
2	27	0	.
3	47	0	3
4	52	.	.

Multiple Imputation with Chained Equations (MICE)

Multiple Imputation with Chained Equations (MICE):[13]

- Imputes data under the (explicit) assumption that the **missingness is MAR**
- Uses a technique known as **chained equations,** which avoids explicit distributional assumptions about the joint data

Chained equations is a method whereby equations are **iteratively estimated.** Specifically, MICE works by (i) specifying an imputation model for each variable with missing data, (ii) starting with a random draw from the observed data, and (iii) using chained equations to iteratively estimate several univariate conditional equations in order to estimate parameters of the conditional densities and impute missing data by drawing from this density.

[13] MICE was developed by van Buuren and Groothuis-Oudshoorn. For more details, see van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1–67.

Multiple Imputation Example

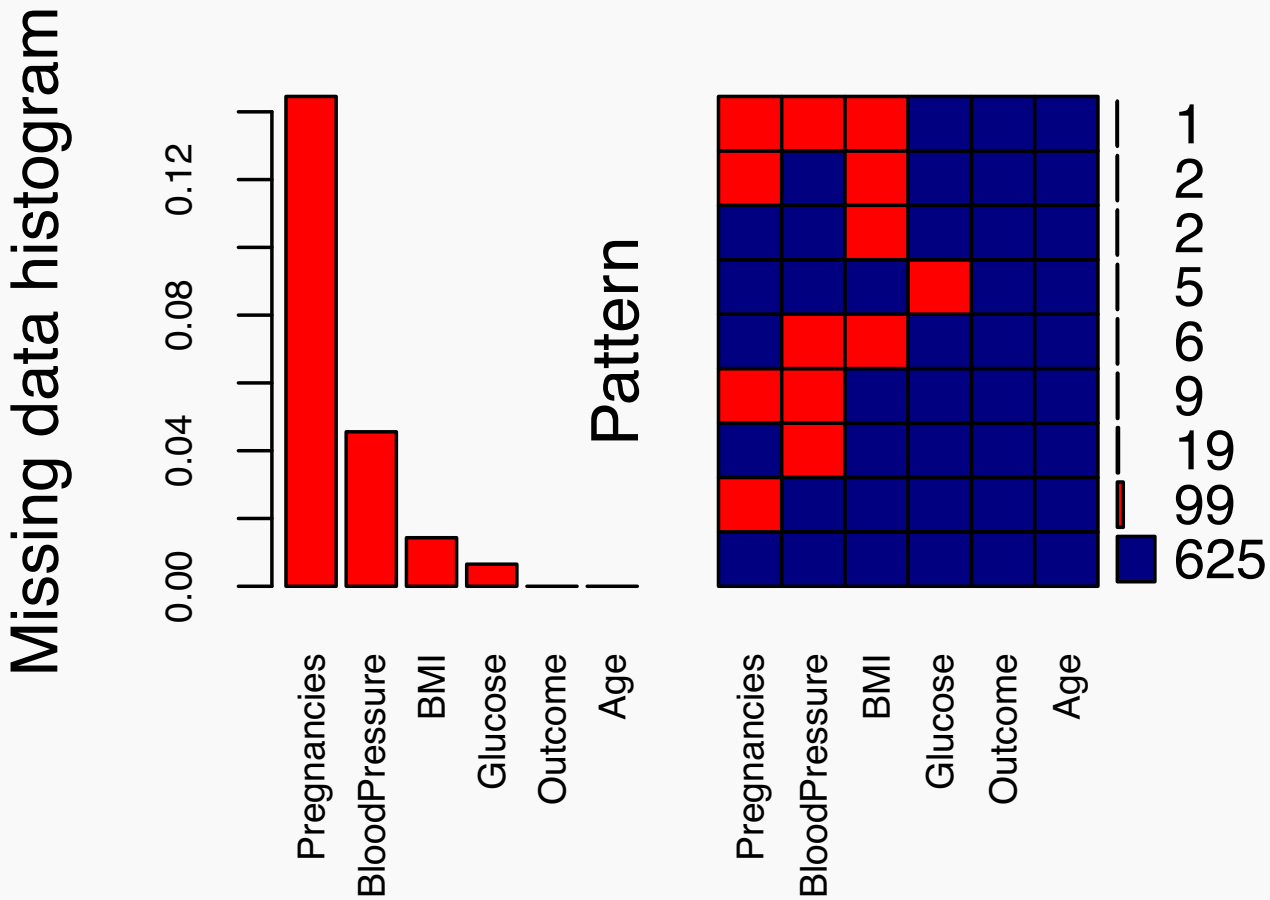
We'll now go through an **example using the *mice* package in R** and a dataset from the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset includes information from **768 individuals aged 21 or older who are of Pima Indian heritage.**

We are interested in estimating the effect of **number of pregnancies** on **risk of diabetes,** adjusting for age.[14]

[14] For the purposes of lab, we will assume this constitutes a well-defined causal question, that data on number of pregnancies are MAR, and that all other assumptions for causal inference are met.

Multiple Imputation: Describing the Pattern of Missing Data

It is helpful to begin by **describing the pattern of missing data.**



Multiple Imputation: 3 Steps

Multiple imputation involves the following **3 steps:**

- **Step 1:** Construct m datasets where $m > 1$. In each of these m datasets, missing values are imputed (i.e., replaced by simulated values) based on a pre-defined procedure.
- **Step 2:** Conduct the analysis of interest across all m datasets.
- **Step 3:** Pool estimates across all m datasets and correctly estimate the variance around these estimates by incorporating information about both within-imputation and between-imputation variability.

We'll go through each of these steps with our pregnancy and diabetes data example.

Multiple Imputation Step 1: Determining m

How many imputed datasets m should we create?

- The efficiency of a result from m imputations relative to an infinite number of imputations is **a function of both m and $\hat{\lambda}$,** where $\hat{\lambda}$ refers to the fraction of missing information about our estimates of interest.
- **A small number of imputations** (m between 5 and 10) is usually sufficient to obtain relatively efficient estimates when using multiple imputation,[15] but recent work suggests that a larger number is usually preferred to maintain statistical power.[16]

For our analysis, computation time is not a concern so we will use $\boldsymbol{m = 40}$ **imputed datasets.**

[15] Rubin, D. B. 1987. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons

[16] Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory, Prev Sci., 2007, vol. 8 3(pg. 206-213)

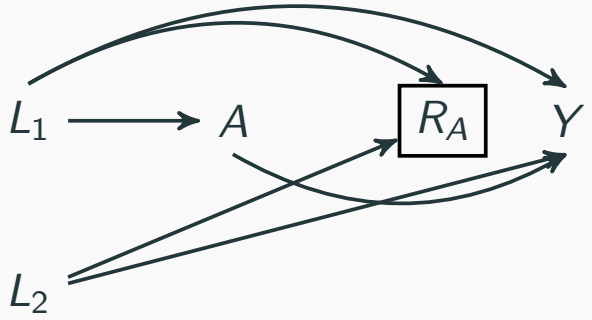
Multiple Imputation Step 1: Variables Included in the Imputation Model

Multiple imputation models should be **at least as rich as the analytic model.** This implies that the multiple imputation model should include:

1. **all variables we will use in our outcome regression model,** and
2. **any other variables which may be important predictors of missing values.**

Multiple Imputation Step 1: Variables Included in the Imputation Model

Recall that we are interested in estimating the effect of number of pregnancies (A) on risk of diabetes (Y), adjusting for age (L_1). So L_1 **must be included in the imputation model.** To impute missing values on pregnancies, we will also use information on the vector of variables L_2 (BMI, blood pressure, and plasma glucose concentration) which we hypothesize are **important predictors of missingness** (and the outcome, in this case):



Multiple Imputation Step 1: Variables Included in the Imputation Model

The **ordering of variables** may also affect results and model convergence. It is often useful to order the variables logically, from least to most missing values.

For example, when the missing data pattern is close to **monotone,** convergence may be fastest when the variables are listed in increasing order based on their proportion of missing values. We can specify this order using the monotone option in the `mice` package:[17]

```
imp <- mice(diabetes, m = 40,
                  maxit = 1,
                  vis =  "monotone",
                  seed = 23390)
```

[17] Check out Stef van Buuren's online book for other useful practical tips.

Multiple Imputation Step 1: Specifying the Imputation Method

MICE provides a number of different **methods to impute the missing values,** including:

- ***norm***: linear regression imputation (numeric data)
- ***pmm***: predictive mean matching (numeric data that isn't normally distributed - the default in the MICE package)
- ***logreg***: logistic regression imputation (binary data)
- ***polyreg***: multinomial/polytomous regression imputation (unordered categorical data)
- ***polr***: ordinal/proportional odds regression imputation (ordered data with $>$ 2 levels)

To see all of the available methods, use the following command in R:

```
methods(mice)
```

Multiple Imputation Step 1: Specifying the Imputation Method

First, we'll **subset to the variables of interest** and order them from least to most percent missing data

```
# Subset to variables of interest, ordered from least to most % missing
diabetes <- subset(diabetes.df, select = c(Outcome, Age, Glucose,
                                           BMI, BloodPressure, Pregnancies))
```

Then we'll use **predictive mean matching** for glucose, BMI, and pregnancies but **linear regression imputation** for blood pressure:

```
library(mice)
impdat <- mice(diabetes, m = 40,
               meth=c("", "",  "pmm", "pmm", "norm", "pmm"),
               seed = 23390)
```

Multiple Imputation Step 2: Conduct the Analysis Across All m Datasets.

We now **perform logistic outcome regression** for the outcome of diabetes, with pregnancies and age as predictors in the model. This analysis is **repeated $m = 40$ times,** once in each of the imputed datasets, using the `with` function:

```
# Run the outcome regression model
mice.results <- with(impdat,
                     glm(Outcome ~ Pregnancies + Age,
                                   family="binomial"))
```

Multiple Imputation Step 3: Pooling Estimates

A **pooled point estimate** can be obtained across the m imputed datasets by taking a simple average:

$$\bar{\beta} = \frac{1}{m} \sum_{k=1}^{m} \widehat{\beta}^{(k)}$$

Rubin's rules can be used to obtain a **pooled variance estimate**, which includes both the within-imputation variance and the between-imputation variance:

$$T = \bar{V}_\beta + (1 + m^{-1})B$$

Where $\bar{V}_\beta = \frac{1}{m} \sum_{k=1}^{m} \widehat{V}_\beta^{(k)}$ is the within-imputation variance, $B = \frac{1}{m-1} \sum_{k=1}^{m} (\widehat{\beta}^{(k)} - \bar{\beta})^2$ is the between-imputation variance, $\widehat{\beta}^{(k)}$ is the point estimate of interest obtained from the k^{th} imputed dataset, and $\widehat{V}_\beta^{(k)}$ is the corresponding variance estimate from the k^{th} imputed dataset for $k = 1, ..., m$.

Multiple Imputation Step 3: Pooling Estimates

Pooling can be accomplished in R using the `mice` package:

```
# Pool estimates and summarize results
summary(pool(mice.results))
```

```
##          term   estimate   std.error statistic       df      p.value
## 1 (Intercept) -2.0159061 0.241655245 -8.342075 759.6363 3.426303e-16
## 2 Pregnancies  0.1299941 0.031141573  4.174295 531.2947 3.492474e-05
## 3         Age  0.0235819 0.007836036  3.009417 726.2923 2.708077e-03
```

Question: What would you expect to happen to the point estimates and standard errors if we performed a naive analysis by pasting together all 40 imputed datasets and analyzing this combined data as if it were a single dataset?

Multiple Imputation: Relative Efficiency Measures

We also defined two useful quantities that help us describe how much missingness is contributing to the **relative inefficiency** of our estimates.

r represents the relative increase in variance due to non-response:

$$r = \frac{(1+m^{-1})B}{\bar{V}_\beta}$$

$\hat{\lambda}$ represents the fraction of missing information:

$$\hat{\lambda} = \frac{r+2/(\nu+3)}{r+1}$$

Where B is once again the between-imputation variance, $\bar{V}_\beta$ is the within-imputation variance, and ν is the degrees of freedom we need to use for our inferences given m, B, and V_β (see lecture slides for the formula for ν). Note that at the limit of m, i.e., when $m \to \infty$, $r = \frac{B}{\bar{V}_\beta}$ and, consequently, $\hat{\lambda} = \frac{B}{T}$, where $T = B + \bar{V}_\beta$.

Ignorability

The notion of **ignorability** brings together ideas connected with both the parameters we are ultimately interested in estimating and the missingness classification. Specifically, missingness is said to be **ignorable** if:[18]

1. Data are **MCAR or MAR;** and
2. Parameters of interest for our analysis (e.g., β) are **distinct** from parameters which govern the distribution of $\boldsymbol{R}$

Missingness is said to be **non-ignorable** if it does not meet the above two criteria. Thus, data that are MNAR are non-ignorable.

[18] For more on ignorability, see Little, R. J. A., and D. B. Rubin. 2002. Statistical Analysis with Missing Data. 2nd ed. New York: John Wiley & Sons (p.119).

Target Validity

A growing body of literature has focused on the idea of **target validity.**[19] Target validity is achieved when **target bias = 0,** i.e., when the total difference between the true causal effect in the target population and the estimated causal effect in the study sample is 0.

The basic idea behind target validity is to provide **a joint measure of the validity of an effect estimate with respect to a specific population of interest** by combining the (previously separate) notions of external and internal validity.

[19] See for example: Westreich D, Edwards JK, Lesko CR, Cole SR, Stuart EA. Target Validity and the Hierarchy of Study Designs. Am J Epidemiol. 2019 Feb 1;188(2):438-443.

Multiple Imputation with Amelia II

We mentioned that multiple imputation can also be performed using **Amelia II,** which is a multiple imputation package developed by Honaker, King, and Blackwell (2019).

This approach imputes missing data under two key assumptions:

1. The complete data, i.e., the data consisting of both non-missing and missing values, follow a **multivariate normal distribution.** Specifically, for a complete dataset Z, $Z \sim \mathcal{N}(\mu, \Sigma)$ where μ refers to a vector of means of all variables in the data and Σ refers to the associated variance-covariance matrix.
2. The **missingness mechanism is MAR.**

Multiple Imputation with Amelia II

To estimate parameters of the multivariate normal distribution, Amelia II uses a procedure called the **Expectation Maximization with Bootstrap (EMB)** algorithm

For a given dataset Z with n observations, Amelia II imputes data in three steps:

1. **Draw a bootstrap sample** of size n with replacement
2. **Run the Expectation Maximization algorithm** in the bootstrap sample to estimate μ and Σ
3. **Use estimates of μ and Σ and the non-missing data** to impute missing observations

Amelia II goes through these three steps m times to create m complete datasets. We can then calculate pooled estimates and variances using Rubin's rules.

How To Choose Between Ameila II vs. MICE?

Some argue that it is better to use **Amelia II** when we have confidence that the joint distribution is indeed **multivariate normal.** However, as the authors of Amelia II argue, even in instances where the multivariate normal distribution is a crude approximation of the true joint distribution, Amelia II "works well".

If you are concerned about the multivariate normal assumption, **MICE** may be an attractive alternative. MICE provides **more flexibility** in modeling the missing variables. MICE may therefore be an attractive procedure if you believe it is **difficult to specify the joint distribution** of the data more generally. However, MICE can lead to the specification of chained equations for which **no joint distribution exists.** The authors of MICE note that some simulation work suggests that this problem may not pose serious threats to imputation quality. Finally, note that **model convergence issues** and **slow processing times** are more common with MICE.