Missing Responses in Longitudinal Data

Example: Study on the Financial Crisis

The data that we analyzed...

id	Age	Sex	Y2	Y3	Y4	Y5	Y6	Y7	Y8
10	23.92	F	2	2	3	2	3	3	3
17	76.09	M	2	4	2	4	4	4	4
29	54.04	F	1	2	2	4	3	3	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1810	46.77	M	2	2	2	2	2	2	2
1826	24.77	M	2	2	2	2	2	3	1
1837	26.29	F	1	1	1	1	1	1	1

Example: Study on the Financial Crisis

The data that were collected...

id	Age	Sex	Y2	Y3	Y4	Y5	Y6	Y7	Y8
2	40.55	F	2	1	3	2	NA	3	NA
10	23.92	F	2	2	3	2	3	3	3
11	44.15	M	NA	NA	NA	NA	1	3	NA
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2439	29.58	M	4	4	NA	NA	2	4	NA
2441	54.64	M	2	1	1	1	1	1	NA
2612	30	M	2	1	1	NA	NA	NA	NA

Missing data refer to any observations which we *intended* to collect, but which were not recorded in our data file for **any reason**.

Missing data is a pervasive problem across most domains, but it is *particularly* common in longitudinal studies.

Classification of Missing Data Mechanisms

Notation for Missingness

We define an **observation indicator** R_{ij}.

$$R_{ij} = \begin{cases} 1 & Y_{ij} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$$

We will assume that we are only concerned with missingness in the **outcomes**, and that R_{ij} is observed for all i, j.

Notation for Missingness

We define an **observation indicator** R_{ij}.

$$R_{ij} = \begin{cases} 1 & Y_{ij} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$$

We will assume that we are only concerned with missingness in the **outcomes**, and that R_{ij} is observed for all i, j.

We partition the **outcome** Y_i into the **observed components** Y_i^{O} and the **missing components**, Y_i^{M}.

A Hierarchy of Types of Missing Data

- Data are said to be **missing completely at random (MCAR)** if

$$f_{R_i}(r_i|Y_i^{\mathrm{O}}, Y_i^{\mathrm{M}}, X_i) = f_{R_i}(r_i|X_i).$$

A Hierarchy of Types of Missing Data

- Data are said to be **missing completely at random (MCAR)** if

$$f_{R_i}(r_i | Y_i^{\mathrm{O}}, Y_i^{\mathrm{M}}, X_i) = f_{R_i}(r_i | X_i).$$

  - Data may be MCAR if, for instance, missingness is due to a hard drive failure on the investigators computer.

A Hierarchy of Types of Missing Data

- Data are said to be **missing completely at random (MCAR)** if

$$f_{R_i}(r_i | Y_i^{\mathrm{O}}, Y_i^{\mathrm{M}}, X_i) = f_{R_i}(r_i | X_i).$$

  - Data may be MCAR if, for instance, missingness is due to a hard drive failure on the investigators computer.

- Data are said to be **missing at random (MAR)** if

$$f_{R_i}(r_i | Y_i^{\mathrm{O}}, Y_i^{\mathrm{M}}, X_i) = f_{R_i}(r_i | Y_i^{\mathrm{O}}, X_i).$$

A Hierarchy of Types of Missing Data

- Data are said to be **missing completely at random (MCAR)** if

$$f_{R_i}(r_i | Y_i^{\mathrm{O}}, Y_i^{\mathrm{M}}, X_i) = f_{R_i}(r_i | X_i).$$

  - Data may be MCAR if, for instance, missingness is due to a hard drive failure on the investigators computer.

- Data are said to be **missing at random (MAR)** if

$$f_{R_i}(r_i | Y_i^{\mathrm{O}}, Y_i^{\mathrm{M}}, X_i) = f_{R_i}(r_i | Y_i^{\mathrm{O}}, X_i).$$

  - Data may be MAR if, for instance, patients who record a severely adverse event ($Y_{ij} > C$ for some known constant C) are removed from the study for all future time points.

A Hierarchy of Types of Missing Data

- Data are said to be **missing completely at random (MCAR)** if

$$f_{R_i}(r_i|Y_i^{\mathrm{O}}, Y_i^{\mathrm{M}}, X_i) = f_{R_i}(r_i|X_i).$$

  - Data may be MCAR if, for instance, missingness is due to a hard drive failure on the investigators computer.

- Data are said to be **missing at random (MAR)** if

$$f_{R_i}(r_i|Y_i^{\mathrm{O}}, Y_i^{\mathrm{M}}, X_i) = f_{R_i}(r_i|Y_i^{\mathrm{O}}, X_i).$$

  - Data may be MAR if, for instance, patients who record a severely adverse event ($Y_{ij} > C$ for some known constant C) are removed from the study for all future time points.

- Otherwise, data are said to be **not missing at random (NMAR)**.

A Hierarchy of Types of Missing Data

- Data are said to be **missing completely at random (MCAR)** if

$$f_{R_i}(r_i | Y_i^{\mathrm{O}}, Y_i^{\mathrm{M}}, X_i) = f_{R_i}(r_i | X_i).$$

  - Data may be MCAR if, for instance, missingness is due to a hard drive failure on the investigators computer.

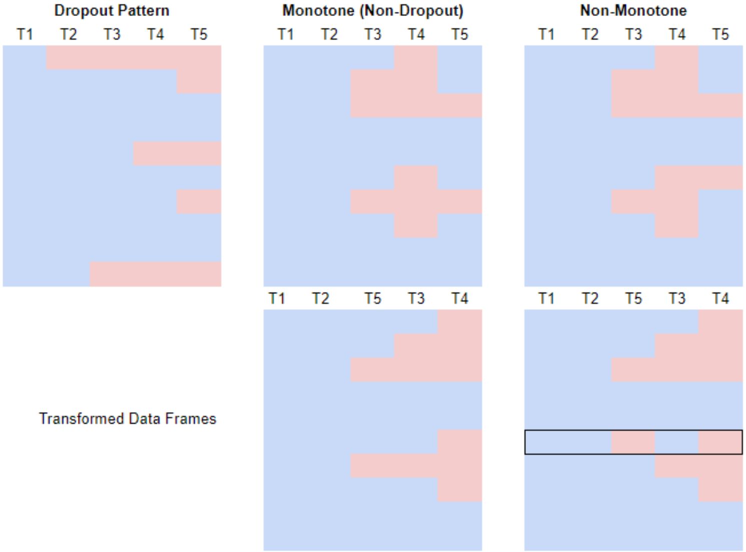
- Data are said to be **missing at random (MAR)** if

$$f_{R_i}(r_i | Y_i^{\mathrm{O}}, Y_i^{\mathrm{M}}, X_i) = f_{R_i}(r_i | Y_i^{\mathrm{O}}, X_i).$$

  - Data may be MAR if, for instance, patients who record a severely adverse event ($Y_{ij} > C$ for some known constant C) are removed from the study for all future time points.

- Otherwise, data are said to be **not missing at random (NMAR)**.
  - Data may be NMAR if, for instance, individuals who smoke more (Y_{ij} large) are less likely to continue responding to a smoking questionnaire.

Patterns of Missingness



Impacts of Missingness

What happens if we ignore missingness?

1. A **complete case** analysis is valid if data are MCAR. However...

What happens if we ignore missingness?

1. A **complete case** analysis is valid if data are MCAR. However...
   - The results will be unnecessarily inefficient and will only be valid if the data are MCAR.

What happens if we ignore missingness?

1. A **complete case** analysis is valid if data are MCAR. However...
   - The results will be unnecessarily inefficient and will only be valid if the data are MCAR.
2. An **available data** analysis is valid if the data are MCAR, and is more efficient than a complete case analysis. However...

What happens if we ignore missingness?

1. A **complete case** analysis is valid if data are MCAR. However...
   - The results will be unnecessarily inefficient and will only be valid if the data are MCAR.
2. An **available data** analysis is valid if the data are MCAR, and is more efficient than a complete case analysis. However...
   - The data will be inherently unbalanced (meaning only certain techniques can be used) and the results will only be valid if the data are MCAR.

What happens if we ignore missingness?

1. A **complete case** analysis is valid if data are MCAR. However...
   - The results will be unnecessarily inefficient and will only be valid if the data are MCAR.
2. An **available data** analysis is valid if the data are MCAR, and is more efficient than a complete case analysis. However...
   - The data will be inherently unbalanced (meaning only certain techniques can be used) and the results will only be valid if the data are MCAR.
3. **Likelihood based techniques** will be valid if the data are MAR or MCAR, so long as the model is **correctly specified**.

What happens if we ignore missingness?

1. A **complete case** analysis is valid if data are MCAR. However...
   - The results will be unnecessarily inefficient and will only be valid if the data are MCAR.
2. An **available data** analysis is valid if the data are MCAR, and is more efficient than a complete case analysis. However...
   - The data will be inherently unbalanced (meaning only certain techniques can be used) and the results will only be valid if the data are MCAR.
3. **Likelihood based techniques** will be valid if the data are MAR or MCAR, so long as the model is **correctly specified**.
4. In all other situations, estimators will be **biased**, and inference will be **invalid**.

General Techniques for Handling Missingness

Families of Techniques

1. **Complete case analysis**, where only the *complete* responders are included in the data frame.

Families of Techniques

1. **Complete case analysis**, where only the *complete* responders are included in the data frame.
2. **Available data analysis**, where all observations that were made are included in the data frame.

Families of Techniques

1. **Complete case analysis**, where only the *complete* responders are included in the data frame.
2. **Available data analysis**, where all observations that were made are included in the data frame.
3. **Weighting techniques**, where pseudo datasets are created based on weighting the available information.

Families of Techniques

1. **Complete case analysis**, where only the *complete* responders are included in the data frame.
2. **Available data analysis**, where all observations that were made are included in the data frame.
3. **Weighting techniques**, where pseudo datasets are created based on weighting the available information.
4. **Imputation techniques**, where the missing values are filled-in based on an underlying model.

A Note on Handling NMAR Missingness

The four classes of techniques listed above will **not** be valid for NMAR data. NMAR data **need** joint modelling strategies for

$$f(Y_i, R_i).$$

Weighting Techniques

Dropout as Missingness

One special type of missingness is **dropout**. In this case, an individual is observed only until t_j, and no times after.

Define D_i to be the **dropout time**

$$D_i = 1 + \sum_{j=1}^{K} R_{ij}.$$

Probability of Inclusion

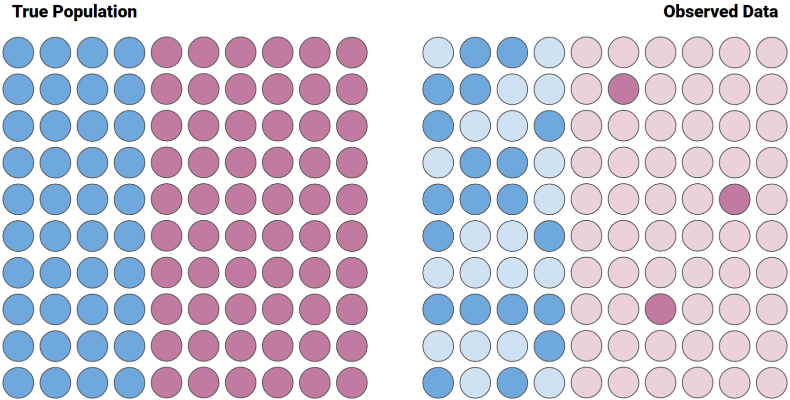
We can think about estimating the **probability of inclusion** for any individual, at any time in the study.

$$\pi_{ij} = P(D_i > j | D_i \geq j, i).$$

This gives the probability that individual i was still under observation at time j, assuming that they made it to at least time $j - 1$. We can estimate these probabilities via (e.g.) logistic regression.

Individuals with **low** π_{ij} were unlikely to have been observed. Individuals with **high** π_{ij} were likely to have been observed.

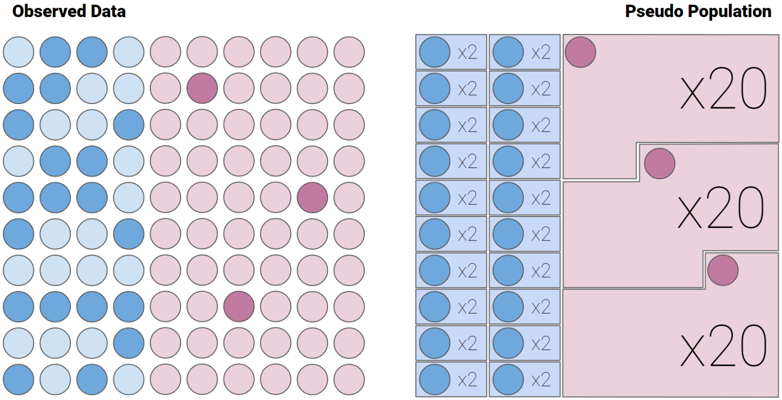
Balancing the Observed Data



$$\pi_B = P(\text{Observed}|\text{Blue}) = \frac{20}{40} = 0.5 \quad \text{and} \quad \pi_M = P(\text{Observed}|\text{Magenta}) = \frac{3}{60} = 0.05.$$

Constructing a Pseudo-Population

$$w_B = \frac{1}{\pi_B} = 2 \quad \text{and} \quad w_M = \frac{1}{\pi_M} = 20.$$



Applying this to Longitudinal Data

We can take

$$\pi_i = P(D_i > K) = \prod_{j=1}^{K} \pi_{ij},$$

and correspondingly get

$$w_i = \frac{1}{\pi_i} = \frac{1}{\prod_{j=1}^{K} \pi_{ij}}.$$

Then, by running a **complete case analysis** with these weights, any MCAR or MAR missingness will be accounted for (assuming the π_i are correct).

More Efficiently

We can extend this idea to weighting **all available data** instead of just the complete cases.

$$w_{ij} = \frac{1}{P(D_i > j)} = \left[\prod_{\ell=1}^{j} \pi_{i\ell} \right]^{-1}.$$

Then giving each individual's observations at time j a weight of w_{ij}, and running an **available data analysis** produces more efficient corrected estimators.

IPW-GEE

Applied specifically to GEE:

- Define $W_i = \text{diag}(R_{ij} w_{ij} \mid j = 1, \ldots, K_i)$.
- Find $\widehat{\beta}$ which solve

$$\sum_{i=1}^{N} D_i' V_i^{-1} W_i \{Y_i - \mu_i(\beta)\} = 0.$$

IPW-GEE

Applied specifically to GEE:

- Define $W_i = \mathrm{diag}(R_{ij} w_{ij} \mid j = 1, \ldots, K_i)$.
- Find $\widehat{\beta}$ which solve
$$\sum_{i=1}^{N} D_i' V_i^{-1} W_i \left\{ Y_i - \mu_i(\beta) \right\} = 0.$$
- We either need X_i fully observed or V_i diagonal.

IPW-GEE

Applied specifically to GEE:

- Define $W_i = \mathrm{diag}(R_{ij} w_{ij} \mid j = 1, \ldots, K_i)$.
- Find $\widehat{\beta}$ which solve

$$\sum_{i=1}^{N} D_i' V_i^{-1} W_i \{Y_i - \mu_i(\beta)\} = 0.$$

- We either need X_i fully observed or V_i diagonal.
  - Recall that V_i need not be correctly specified. If X_i is not fully observed, take V_i diagonal.

Imputation Techniques

Imputation in General

- Using some model, **estimate the missing values** Y_i^{M} based on the observed values, Y_i^{O} and variates X_i.

Imputation in General

- Using some model, **estimate the missing values** Y_i^{M} based on the observed values, Y_i^{O} and variates X_i.
- **Compute the parameters** of interest as though these imputed values were the truth.

Imputation in General

- Using some model, **estimate the missing values** Y_i^{M} based on the observed values, Y_i^{O} and variates X_i.
- **Compute the parameters** of interest as though these imputed values were the truth.
- If using **multiple imputation** repeat this procedure m times, averaging the results.

Imputation in General

- Using some model, **estimate the missing values** Y_i^{M} based on the observed values, Y_i^{O} and variates X_i.
- **Compute the parameters** of interest as though these imputed values were the truth.
- If using **multiple imputation** repeat this procedure m times, averaging the results.

Need to choose **single** or **multiple** imputation, and the **imputation procedure**.

Multiple Imputation

Repeat the imputation process m times, giving $\widehat{\beta}^{(k)}$ for $k = 1, \ldots, m$. Then

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

and we further take

$$\widehat{\text{cov}}\left(\widehat{\beta}\right) = \frac{1}{m} \sum_{k=1}^{m} \text{cov}\left(\widehat{\beta}^{(k)}\right) + \frac{m+1}{m(m-1)} \sum_{k=1}^{m} \left(\widehat{\beta}^{(k)} - \widehat{\beta}\right)\left(\widehat{\beta}^{(k)} - \widehat{\beta}\right)'.$$

Regression Based Imputation Procedure

1. Fit a GLM $g(E[Y_{i2}|Y_{i1}, X_i]) = Z_{i1}'\gamma_2$.

Regression Based Imputation Procedure

1. Fit a GLM $g(E[Y_{i2}|Y_{i1}, X_i]) = Z_{i1}'\gamma_2$.
2. This gives **predictions**, $\widehat{Y}_{i2} = g^{-1}(Z_{i1}'\widehat{\gamma}_2)$ for all with $R_{i2} = 0$.

Regression Based Imputation Procedure

1. Fit a GLM $g(E[Y_{i2}|Y_{i1}, X_i]) = Z_{i1}'\gamma_2$.
2. This gives **predictions**, $\widehat{Y}_{i2} = g^{-1}(Z_{i1}'\widehat{\gamma}_2)$ for all with $R_{i2} = 0$.
3. To ensure that predictions are **not deterministic** consider **sampling from** the distribution with mean $\widehat{Y}_{i2}$.

Regression Based Imputation Procedure

1. Fit a GLM $g(E[Y_{i2}|Y_{i1}, X_i]) = Z_{i1}'\gamma_2$.
2. This gives **predictions**, $\widehat{Y}_{i2} = g^{-1}(Z_{i1}'\widehat{\gamma}_2)$ for all with $R_{i2} = 0$.
3. To ensure that predictions are **not deterministic** consider **sampling from** the distribution with mean $\widehat{Y}_{i2}$.
4. Repeat the process for $g(E[Y_{i3}|Y_{i1}, Y_{i2}, X_i]) = Z_{i2}'\gamma_3$, where the **sampled values** replace any missing values.

Regression Based Imputation Procedure

1. Fit a GLM $g(E[Y_{i2}|Y_{i1}, X_i]) = Z_{i1}'\gamma_2$.
2. This gives **predictions**, $\widehat{Y}_{i2} = g^{-1}(Z_{i1}'\widehat{\gamma}_2)$ for all with $R_{i2} = 0$.
3. To ensure that predictions are **not deterministic** consider **sampling from** the distribution with mean $\widehat{Y}_{i2}$.
4. Repeat the process for $g(E[Y_{i3}|Y_{i1}, Y_{i2}, X_i]) = Z_{i2}'\gamma_3$, where the **sampled values** replace any missing values.
5. Continue imputing, estimate $\widehat{\beta}^{(1)}$, and then repeat m times.

Problem with Underestimating Uncertainty

This procedure outlined **underestimates** the variability that should be inherent to this imputation procedure, since $\widehat{Y}_{ij}$ is estimated not fixed!

Instead of using estimated $\widehat{\gamma}_j$, we **draw from the posterior distribution**, giving $\widetilde{\gamma}_j$, and otherwise proceed as outlined.

Predictive Mean Matching

Instead of using the regression models to **sample predicted values**, we can use them to **match** individuals who have observations recorded at the relevant time.

1. Generate $\widehat{Y}_{ij}$ for all i.

Predictive Mean Matching

Instead of using the regression models to **sample predicted values**, we can use them to **match** individuals who have observations recorded at the relevant time.

1. Generate $\widehat{Y}_{ij}$ for all i.
2. For each Y_{ij}^{M}, select the κ nearest individuals with Y_{ij}^{O}, based on $\widehat{Y}_{ij}$.

Predictive Mean Matching

Instead of using the regression models to **sample predicted values**, we can use them to **match** individuals who have observations recorded at the relevant time.

1. Generate $\widehat{Y}_{ij}$ for all i.
2. For each Y_{ij}^{M}, select the κ nearest individuals with Y_{ij}^{O}, based on $\widehat{Y}_{ij}$.
3. Sample one of these κ, and use the observed Y_{ij} as the value.

Predictive Mean Matching

Instead of using the regression models to **sample predicted values**, we can use them to **match** individuals who have observations recorded at the relevant time.

1. Generate $\widehat{Y}_{ij}$ for all i.
2. For each Y_{ij}^{M}, select the κ nearest individuals with Y_{ij}^{O}, based on $\widehat{Y}_{ij}$.
3. Sample one of these κ, and use the observed Y_{ij} as the value.
4. Repeat this process for all j, and then m times.

Likelihood as Imputation

Any technique that uses **maximum likelihood** (e.g., GLMEMs or transition models) will result in valid inference if the data are **MCAR** or **MAR**. In this case

$$f(Y_i|X_i) = f(Y_i^{\mathrm{O}}|X_i) = f(Y_i^{\mathrm{M}}|X_i).$$

There are procedures (using Expectation Maximization (EM)) which make the connection between **likelihood** and **imputation** more clear.

Summary

- Missing data is a **pervasive issue** in longitudinal studies.

Summary

- Missing data is a **pervasive issue** in longitudinal studies.
- Ignoring missingness causes **loss of efficiency** in the best case, and can **completely invalidate** analyses in the worst.

Summary

- Missing data is a **pervasive issue** in longitudinal studies.
- Ignoring missingness causes **loss of efficiency** in the best case, and can **completely invalidate** analyses in the worst.
- Missingness is categorized as **MCAR**, **MAR**, or **NMAR**, based on how it relates to the observed data.

Summary

- Missing data is a **pervasive issue** in longitudinal studies.
- Ignoring missingness causes **loss of efficiency** in the best case, and can **completely invalidate** analyses in the worst.
- Missingness is categorized as **MCAR**, **MAR**, or **NMAR**, based on how it relates to the observed data.
- Missing is easier to handle when it is **monotone** (for instance, based on **dropout**).

Summary

- Missing data is a **pervasive issue** in longitudinal studies.
- Ignoring missingness causes **loss of efficiency** in the best case, and can **completely invalidate** analyses in the worst.
- Missingness is categorized as **MCAR**, **MAR**, or **NMAR**, based on how it relates to the observed data.
- Missing is easier to handle when it is **monotone** (for instance, based on **dropout**).
- **Complete case analyses** and **available data analyses** provide valid inference only under MCAR.

Summary

- Missing data is a **pervasive issue** in longitudinal studies.
- Ignoring missingness causes **loss of efficiency** in the best case, and can **completely invalidate** analyses in the worst.
- Missingness is categorized as **MCAR**, **MAR**, or **NMAR**, based on how it relates to the observed data.
- Missing is easier to handle when it is **monotone** (for instance, based on **dropout**).
- **Complete case analyses** and **available data analyses** provide valid inference only under MCAR.
- **Weighting techniques** generate pseudo-populations that match the would-be observed population using estimated probabilities.

Summary

- Missing data is a **pervasive issue** in longitudinal studies.
- Ignoring missingness causes **loss of efficiency** in the best case, and can **completely invalidate** analyses in the worst.
- Missingness is categorized as **MCAR**, **MAR**, or **NMAR**, based on how it relates to the observed data.
- Missing is easier to handle when it is **monotone** (for instance, based on **dropout**).
- **Complete case analyses** and **available data analyses** provide valid inference only under MCAR.
- **Weighting techniques** generate pseudo-populations that match the would-be observed population using estimated probabilities.
- **Imputation techniques** fill in the missing values based on specific regression models.