# Introduction to Handling Missing Data (in R)

Dan Hall, Director of the SCC



Department of Statistics
*Franklin College of Arts and Sciences*
*Statistical Consulting Center*
UNIVERSITY OF GEORGIA

# Table of Contents

The Problem of Missing Data

Missing Data Typology

Missing Data Approaches
- Complete Case Analysis (aka Listwise Deletion).
- Available Case Analysis
- (Single) Imputation
- Multiple Imputation
- Full Information Maximum Likelihood Estimation (FIML)

Recommendations

Final Comments

References & Resources

Questions?

# Related Resources

- An R Markdown document in HTML format called `MissingDataNotes.html` accompanies these slides and will be posted online as an attachment to the recording of this seminar.
- The talk will be given from this document as well as from `Missing DataNotes.html`, the document mentioned above.
- The file `Missing DataNotes.html` contains more material, including examples, than we can cover today, so I will post a companion video discussing that material.
- Look for these videos at `https://www.stat.uga.edu/events/past/scc-seminar-series-on-data-analysis`

# Related Resources

- An R Markdown document in HTML format called `MissingDataNotes.html` accompanies these slides and will be posted online as an attachment to the recording of this seminar.
- The talk will be given from this document as well as from `MissingDataNotes.html`, the document mentioned above.
- The file `MissingDataNotes.html` contains more material, including examples, than we can cover today, so I will post a companion video discussing that material.
- Look for these videos at `https://www.stat.uga.edu/events/past/scc-seminar-series-on-data-analysis`

# Related Resources

- An R Markdown document in HTML format called `MissingDataNotes.html` accompanies these slides and will be posted online as an attachment to the recording of this seminar.
- The talk will be given from this document as well as from `Missing DataNotes.html`, the document mentioned above.
- The file `Missing DataNotes.html` contains more material, including examples, than we can cover today, so I will post a companion video discussing that material.
- Look for these videos at `https://www.stat.uga.edu/events/past/scc-seminar-series-on-data-analysis`

# Related Resources

- An R Markdown document in HTML format called `MissingDataNotes.html` accompanies these slides and will be posted online as an attachment to the recording of this seminar.
- The talk will be given from this document as well as from `Missing DataNotes.html`, the document mentioned above.
- The file `Missing DataNotes.html` contains more material, including examples, than we can cover today, so I will post a companion video discussing that material.
- Look for these videos at `https://www.stat.uga.edu/events/past/scc-seminar-series-on-data-analysis`

# The Problem of Missing Data

- Missing data occur when data that we intended to collect were not obtained.
- This can happen in the response, the explanatory variables, or both.
- We assume data on some variables were obtained, excluding complete non-response.
- Longitudinal data: Attrition (loss to follow-up), intermittent missingness (missed appointments).
- Surveys: item skipping, failure to complete.
- Technical problems, lost IDs, sensor glitches, poor records.
- Distinct from censoring and missingness by design or skip logic (structural missingness).
- Depending on the nature of the missingness, ignoring it can cause (potentially severe) bias, inefficiency, or both; can also make some analyses infeasible.

# The Problem of Missing Data

- Missing data occur when data that we intended to collect were not obtained.
- This can happen in the response, the explanatory variables, or both.
- We assume data on some variables were obtained, excluding complete non-response.
- Longitudinal data: Attrition (loss to follow-up), intermittent missingness (missed appointments).
- Surveys: item skipping, failure to complete.
- Technical problems, lost IDs, sensor glitches, poor records.
- Distinct from censoring and missingness by design or skip logic (structural missingness).
- Depending on the nature of the missingness, ignoring it can cause (potentially severe) bias, inefficiency, or both; can also make some analyses infeasible.

# The Problem of Missing Data

- Missing data occur when data that we intended to collect were not obtained.
- This can happen in the response, the explanatory variables, or both.
- We assume data on some variables were obtained, excluding complete non-response.
- Longitudinal data: Attrition (loss to follow-up), intermittent missingness (missed appointments).
- Surveys: item skipping, failure to complete.
- Technical problems, lost IDs, sensor glitches, poor records.
- Distinct from censoring and missingness by design or skip logic (structural missingness).
- Depending on the nature of the missingness, ignoring it can cause (potentially severe) bias, inefficiency, or both; can also make some analyses infeasible.

# The Problem of Missing Data

- Missing data occur when data that we intended to collect were not obtained.
- This can happen in the response, the explanatory variables, or both.
- We assume data on some variables were obtained, excluding complete non-response.
- Longitudinal data: Attrition (loss to follow-up), intermittent missingness (missed appointments).
- Surveys: item skipping, failure to complete.
- Technical problems, lost IDs, sensor glitches, poor records.
- Distinct from censoring and missingness by design or skip logic (structural missingness).
- Depending on the nature of the missingness, ignoring it can cause (potentially severe) bias, inefficiency, or both; can also make some analyses infeasible.

# The Problem of Missing Data

- Missing data occur when data that we intended to collect were not obtained.
- This can happen in the response, the explanatory variables, or both.
- We assume data on some variables were obtained, excluding complete non-response.
- Longitudinal data: Attrition (loss to follow-up), intermittent missingness (missed appointments).
- Surveys: item skipping, failure to complete.
- Technical problems, lost IDs, sensor glitches, poor records.
- Distinct from censoring and missingness by design or skip logic (structural missingness).
- Depending on the nature of the missingness, ignoring it can cause (potentially severe) bias, inefficiency, or both; can also make some analyses infeasible.

# The Problem of Missing Data

- Missing data occur when data that we intended to collect were not obtained.
- This can happen in the response, the explanatory variables, or both.
- We assume data on some variables were obtained, excluding complete non-response.
- Longitudinal data: Attrition (loss to follow-up), intermittent missingness (missed appointments).
- Surveys: item skipping, failure to complete.
- Technical problems, lost IDs, sensor glitches, poor records.
- Distinct from censoring and missingness by design or skip logic (structural missingness).
- Depending on the nature of the missingness, ignoring it can cause (potentially severe) bias, inefficiency, or both; can also make some analyses infeasible.

# The Problem of Missing Data

- Missing data occur when data that we intended to collect were not obtained.
- This can happen in the response, the explanatory variables, or both.
- We assume data on some variables were obtained, excluding complete non-response.
- Longitudinal data: Attrition (loss to follow-up), intermittent missingness (missed appointments).
- Surveys: item skipping, failure to complete.
- Technical problems, lost IDs, sensor glitches, poor records.
- Distinct from censoring and missingness by design or skip logic (structural missingness).
- Depending on the nature of the missingness, ignoring it can cause (potentially severe) bias, inefficiency, or both; can also make some analyses infeasible.

# The Problem of Missing Data

- Missing data occur when data that we intended to collect were not obtained.
- This can happen in the response, the explanatory variables, or both.
- We assume data on some variables were obtained, excluding complete non-response.
- Longitudinal data: Attrition (loss to follow-up), intermittent missingness (missed appointments).
- Surveys: item skipping, failure to complete.
- Technical problems, lost IDs, sensor glitches, poor records.
- Distinct from censoring and missingness by design or skip logic (structural missingness).
- Depending on the nature of the missingness, ignoring it can cause (potentially severe) bias, inefficiency, or both; can also make some analyses infeasible.

# Missing Data Typology

The appropriate way to handle missing data depends upon how the missingness occurs. A widely used categorization distinguishes three types:

1. Missing completely at random (MCAR)
   - Missingness is independent of the unobserved response and of the covariates.
   - E.g., observation logs got wet making some data ilegible.
2. Missing at random (MAR)
   - Missingness is MAR if it is MCAR or if it depends on observed data, but not on the unobserved response.
   - E.g., missingness differs by sex, which is known; or in a weight loss study with weekly weigh-ins, participants drop out after they've met their weight loss goal (missingness depends on past observed value of the response).
3. Not missing at random (NMAR, aka MNAR)
   - Missingness depends on the unobserved response.
   - E.g., in the weight loss study, suppose subjects are more likely to attend weigh-ins when the've met their weekly weight loss goal.

# Missing Data Typology

The appropriate way to handle missing data depends upon how the missingness occurs. A widely used categorization distinguishes three types:

1. Missing completely at random (MCAR)
   - Missingness is independent of the unobserved response and of the covariates.
   - E.g., observation logs got wet making some data ilegible.
2. Missing at random (MAR)
   - Missingness is MAR if it is MCAR or if it depends on observed data, but not on the unobserved response.
   - E.g., missingness differs by sex, which is known; or in a weight loss study with weekly weigh-ins, participants drop out after they've met their weight loss goal (missingness depends on past observed value of the response).
3. Not missing at random (NMAR, aka MNAR)
   - Missingness depends on the unobserved response.
   - E.g., in the weight loss study, suppose subjects are more likely to attend weigh-ins when the've met their weekly weight loss goal.

# Missing Data Typology

The appropriate way to handle missing data depends upon how the missingness occurs. A widely used categorization distinguishes three types:

1. Missing completely at random (MCAR)
   - Missingness is independent of the unobserved response and of the covariates.
   - E.g., observation logs got wet making some data ilegible.
2. Missing at random (MAR)
   - Missingness is MAR if it is MCAR or if it depends on observed data, but not on the unobserved response.
   - E.g., missingness differs by sex, which is known; or in a weight loss study with weekly weigh-ins, participants drop out after they've met their weight loss goal (missingness depends on past observed value of the response).
3. Not missing at random (NMAR, aka MNAR)
   - Missingness depends on the unobserved response.
   - E.g., in the weight loss study, suppose subjects are more likely to attend weigh-ins when the've met their weekly weight loss goal.

# Missing Data Typology

The appropriate way to handle missing data depends upon how the missingness occurs. A widely used categorization distinguishes three types:

1. Missing completely at random (MCAR)
   - Missingness is independent of the unobserved response and of the covariates.
   - E.g., observation logs got wet making some data ilegible.
2. Missing at random (MAR)
   - Missingness is MAR if it is MCAR or if it depends on observed data, but not on the unobserved response.
   - E.g., missingness differs by sex, which is known; or in a weight loss study with weekly weigh-ins, participants drop out after they've met their weight loss goal (missingness depends on past observed value of the response).
3. Not missing at random (NMAR, aka MNAR)
   - Missingness depends on the unobserved response.
   - E.g., in the weight loss study, suppose subjects are more likely to attend weigh-ins when the've met their weekly weight loss goal.

# Missing Data Typology

The appropriate way to handle missing data depends upon how the missingness occurs. A widely used categorization distinguishes three types:

1. Missing completely at random (MCAR)
   - Missingness is independent of the unobserved response and of the covariates.
   - E.g., observation logs got wet making some data ilegible.
2. Missing at random (MAR)
   - Missingness is MAR if it is MCAR or if it depends on observed data, but not on the unobserved response.
   - E.g., missingness differs by sex, which is known; or in a weight loss study with weekly weigh-ins, participants drop out after they've met their weight loss goal (missingness depends on past observed value of the response).
3. Not missing at random (NMAR, aka MNAR)
   - Missingness depends on the unobserved response.
   - E.g., in the weight loss study, suppose subjects are more likely to attend weigh-ins when the've met their weekly weight loss goal.

# Missing Data Typology

The appropriate way to handle missing data depends upon how the missingness occurs. A widely used categorization distinguishes three types:

1. Missing completely at random (MCAR)
   - Missingness is independent of the unobserved response and of the covariates.
   - E.g., observation logs got wet making some data ilegible.
2. Missing at random (MAR)
   - Missingness is MAR if it is MCAR or if it depends on observed data, but not on the unobserved response.
   - E.g., missingness differs by sex, which is known; or in a weight loss study with weekly weigh-ins, participants drop out after they've met their weight loss goal (missingness depends on past observed value of the response).
3. Not missing at random (NMAR, aka MNAR)
   - Missingness depends on the unobserved response.
   - E.g., in the weight loss study, suppose subjects are more likely to attend weigh-ins when the've met their weekly weight loss goal.

# Missing Data Typology

The appropriate way to handle missing data depends upon how the missingness occurs. A widely used categorization distinguishes three types:

1. Missing completely at random (MCAR)
   - Missingness is independent of the unobserved response and of the covariates.
   - E.g., observation logs got wet making some data ilegible.
2. Missing at random (MAR)
   - Missingness is MAR if it is MCAR or if it depends on observed data, but not on the unobserved response.
   - E.g., missingness differs by sex, which is known; or in a weight loss study with weekly weigh-ins, participants drop out after they've met their weight loss goal (missingness depends on past observed value of the response).
3. Not missing at random (NMAR, aka MNAR)
   - Missingness depends on the unobserved response.
   - E.g., in the weight loss study, suppose subjects are more likely to attend weigh-ins when the've met their weekly weight loss goal.

# Missing Data Typology

The appropriate way to handle missing data depends upon how the missingness occurs. A widely used categorization distinguishes three types:

1. Missing completely at random (MCAR)
   - Missingness is independent of the unobserved response and of the covariates.
   - E.g., observation logs got wet making some data ilegible.
2. Missing at random (MAR)
   - Missingness is MAR if it is MCAR or if it depends on observed data, but not on the unobserved response.
   - E.g., missingness differs by sex, which is known; or in a weight loss study with weekly weigh-ins, participants drop out after they've met their weight loss goal (missingness depends on past observed value of the response).
3. Not missing at random (NMAR, aka MNAR)
   - Missingness depends on the unobserved response.
   - E.g., in the weight loss study, suppose subjects are more likely to attend weigh-ins when the've met their weekly weight loss goal.

# Missing Data Typology

The appropriate way to handle missing data depends upon how the missingness occurs. A widely used categorization distinguishes three types:

1. Missing completely at random (MCAR)
   - Missingness is independent of the unobserved response and of the covariates.
   - E.g., observation logs got wet making some data ilegible.
2. Missing at random (MAR)
   - Missingness is MAR if it is MCAR or if it depends on observed data, but not on the unobserved response.
   - E.g., missingness differs by sex, which is known; or in a weight loss study with weekly weigh-ins, participants drop out after they've met their weight loss goal (missingness depends on past observed value of the response).
3. Not missing at random (NMAR, aka MNAR)
   - Missingness depends on the unobserved response.
   - E.g., in the weight loss study, suppose subjects are more likely to attend weigh-ins when the've met their weekly weight loss goal.

# Ignorable Missingness

The term **ignorable** missingness roughly equates to MAR. But this is slightly simplistic.

- To be more precise, ignorable missingness refers to situations when a correctly specified likelihood-based analysis is appropriate. A sufficient condition for ignorability is when
  (a) data are MAR, and
  (b) the models for the observed data and the missingness mechanism do not share parameters.
- When missingness is NMAR, the missingness is non-ignorable.

# Ignorable Missingness

The term **ignorable** missingness roughly equates to MAR. But this is slightly simplistic.

- To be more precise, ignorable missingness refers to situations when a correctly specified likelihood-based analysis is appropriate. A sufficient condition for ignorability is when
  (a) data are MAR, and
  (b) the models for the observed data and the missingness mechanism do not share parameters.
- When missingness is NMAR, the missingness is non-ignorable.

# Ignorable Missingness

The term **ignorable** missingness roughly equates to MAR. But this is slightly simplistic.

- To be more precise, ignorable missingness refers to situations when a correctly specified likelihood-based analysis is appropriate. A sufficient condition for ignorability is when
  (a) data are MAR, and
  (b) the models for the observed data and the missingness mechanism do not share parameters.
- When missingness is NMAR, the missingness is non-ignorable.

# Ignorable Missingness

The term **ignorable** missingness roughly equates to MAR. But this is slightly simplistic.

- To be more precise, ignorable missingness refers to situations when a correctly specified likelihood-based analysis is appropriate. A sufficient condition for ignorability is when
  (a) data are MAR, and
  (b) the models for the observed data and the missingness mechanism do not share parameters.
- When missingness is NMAR, the missingness is non-ignorable.

# Complete Case Analysis (aka Listwise Deletion).

- Means deleting any observation that has a missing value on any variable ($Y$'s and $X$'s)
- Default in much statistical software.
- Inefficient and can cause bias if data are not MCAR.

# Complete Case Analysis (aka Listwise Deletion).

- Means deleting any observation that has a missing value on any variable ($Y$'s and $X$'s)
- Default in much statistical software.
- Inefficient and can cause bias if data are not MCAR.

# Complete Case Analysis (aka Listwise Deletion).

- Means deleting any observation that has a missing value on any variable ($Y$'s and $X$'s)
- Default in much statistical software.
- Inefficient and can cause bias if data are not MCAR.

# Available Case Analysis

- Sometimes, for an observation with missing values on some variables, other variables can still contribute to the analysis. Thus, the entire observation need not be deleted.
- More efficient than complete case analysis, but not necessarily fully efficient or unbiased.
- See examples in MissingDataNotes.html.

# Available Case Analysis

- Sometimes, for an observation with missing values on some variables, other variables can still contribute to the analysis. Thus, the entire observation need not be deleted.
- More efficient than complete case analysis, but not necessarily fully efficient or unbiased.
- See examples in MissingDataNotes.html.

# Available Case Analysis

- Sometimes, for an observation with missing values on some variables, other variables can still contribute to the analysis. Thus, the entire observation need not be deleted.
- More efficient than complete case analysis, but not necessarily fully efficient or unbiased.
- See examples in MissingDataNotes.html.

# (Single) Imputation

There are many ways to fill in, or impute, missing values. Some of these are deterministic, some are stochastic.

- Deterministic methods:
  - *Mean imputation* replaces missing values with that variable's mean of non-missing data. For categorical variables, the modal may be used.
  - *Hot deck* methods replace a missing value with the value from a "similar" observation. E.g., sort the data according to similarity and use the "last-value-carried-forward".
  - *Regression imputation* fits a regression model to the variable with missing data and uses the predicted values to impute missing data.
- Stochastic methods. Bad deterministic methods can introduce location bias. But even good deterministic methods produce data with reduced variability. An appealing solution is to add variability.
  - E.g., regression imputation can be altered by adding error to the predicted value from the fitted regression model. E.g., add a randomly generated error term that is $N[0, s^2]$, where $s$ =RMSE from the fitted model.

# (Single) Imputation

There are many ways to fill in, or impute, missing values. Some of these are deterministic, some are stochastic.

- Deterministic methods:
  - *Mean imputation* replaces missing values with that variable's mean of non-missing data. For categorical variables, the modal may be used.
  - *Hot deck* methods replace a missing value with the value from a "similar" observation. E.g., sort the data according to similarity and use the "last-value-carried-forward".
  - *Regression imputation* fits a regression model to the variable with missing data and uses the predicted values to impute missing data.
- Stochastic methods. Bad deterministic methods can introduce location bias. But even good deterministic methods produce data with reduced variability. An appealing solution is to add variability.
  - E.g., regression imputation can be altered by adding error to the predicted value from the fitted regression model. E.g., add a randomly generated error term that is $N[0, s^2]$, where $s$ =RMSE from the fitted model.

# (Single) Imputation

There are many ways to fill in, or impute, missing values. Some of these are deterministic, some are stochastic.

- Deterministic methods:
  - *Mean imputation* replaces missing values with that variable's mean of non-missing data. For categorical variables, the modal may be used.
  - *Hot deck* methods replace a missing value with the value from a "similar" observation. E.g., sort the data according to similarity and use the "last-value-carried-forward".
  - *Regression imputation* fits a regression model to the variable with missing data and uses the predicted values to impute missing data.
- Stochastic methods. Bad deterministic methods can introduce location bias. But even good deterministic methods produce data with reduced variability. An appealing solution is to add variability.
  - E.g., regression imputation can be altered by adding error to the predicted value from the fitted regression model. E.g., add a randomly generated error term that is $N[0, s^2]$, where $s$ =RMSE from the fitted model.

# (Single) Imputation

There are many ways to fill in, or impute, missing values. Some of these are deterministic, some are stochastic.

- Deterministic methods:
  - *Mean imputation* replaces missing values with that variable's mean of non-missing data. For categorical variables, the modal may be used.
  - *Hot deck* methods replace a missing value with the value from a "similar" observation. E.g., sort the data according to similarity and use the "last-value-carried-forward".
  - *Regression imputation* fits a regression model to the variable with missing data and uses the predicted values to impute missing data.
- Stochastic methods. Bad deterministic methods can introduce location bias. But even good deterministic methods produce data with reduced variability. An appealing solution is to add variability.
  - E.g., regression imputation can be altered by adding error to the predicted value from the fitted regression model. E.g., add a randomly generated error term that is $N[0, s^2]$, where $s$ =RMSE from the fitted model.

# (Single) Imputation

There are many ways to fill in, or impute, missing values. Some of these are deterministic, some are stochastic.

- Deterministic methods:
  - *Mean imputation* replaces missing values with that variable's mean of non-missing data. For categorical variables, the modal may be used.
  - *Hot deck* methods replace a missing value with the value from a "similar" observation. E.g., sort the data according to similarity and use the "last-value-carried-forward".
  - *Regression imputation* fits a regression model to the variable with missing data and uses the predicted values to impute missing data.
- Stochastic methods. Bad deterministic methods can introduce location bias. But even good deterministic methods produce data with reduced variability. An appealing solution is to add variability.
  - E.g., regression imputation can be altered by adding error to the predicted value from the fitted regression model. E.g., add a randomly generated error term that is $N(0, s^2)$, where $s$ =RMSE from the fitted model.

# (Single) Imputation

There are many ways to fill in, or impute, missing values. Some of these are deterministic, some are stochastic.

- Deterministic methods:
  - *Mean imputation* replaces missing values with that variable's mean of non-missing data. For categorical variables, the modal may be used.
  - *Hot deck* methods replace a missing value with the value from a "similar" observation. E.g., sort the data according to similarity and use the "last-value-carried-forward".
  - *Regression imputation* fits a regression model to the variable with missing data and uses the predicted values to impute missing data.
- Stochastic methods. Bad deterministic methods can introduce location bias. But even good deterministic methods produce data with reduced variability. An appealing solution is to add variability.
  - E.g., regression imputation can be altered by adding error to the predicted value from the fitted regression model. E.g., add a randomly generated error term that is $N(0, s^2)$, where $s$ =RMSE from the fitted model.

# Multiple Imputation

- Although stochastic imputation methods diminish the underestimation of variability in the data, *they do not eliminate it.*
- Therefore, inferences from singly imputed data sets are usually liberal:
  - SEs are underestimated,
  - test statistics are too large,
  - $p$-values are too small,
  - confidence intervals are too tight.

A solution to the underestimation of variability associated with single imputation is to do it repeatedly, capturing the variability from one imputation to another. Such an approach is called *multiple imputation* or MI.

- In R, the most popular package for MI is `mice`.

# Multiple Imputation

- Although stochastic imputation methods diminish the underestimation of variability in the data, *they do not eliminate it.*
- Therefore, inferences from singly imputed data sets are usually liberal:
  - SEs are underestimated,
  - test statistics are too large,
  - $p$-values are too small,
  - confidence intervals are too tight.

A solution to the underestimation of variability associated with single imputation is to do it repeatedly, capturing the variability from one imputation to another. Such an approach is called *multiple imputation* or MI.

- In R, the most popular package for MI is `mice`.

# Multiple Imputation

- Although stochastic imputation methods diminish the underestimation of variability in the data, *they do not eliminate it.*
- Therefore, inferences from singly imputed data sets are usually liberal:
  - SEs are underestimated,
  - test statistics are too large,
  - $p$-values are too small,
  - confidence intervals are too tight.

A solution to the underestimation of variability associated with single imputation is to do it repeatedly, capturing the variability from one imputation to another. Such an approach is called *multiple imputation* or MI.

- In R, the most popular package for MI is `mice`.

# Multiple Imputation

- Although stochastic imputation methods diminish the underestimation of variability in the data, *they do not eliminate it.*
- Therefore, inferences from singly imputed data sets are usually liberal:
  - SEs are underestimated,
  - test statistics are too large,
  - $p$-values are too small,
  - confidence intervals are too tight.

A solution to the underestimation of variability associated with single imputation is to do it repeatedly, capturing the variability from one imputation to another. Such an approach is called *multiple imputation* or MI.

- In R, the most popular package for MI is `mice`.

# Multiple Imputation

- Although stochastic imputation methods diminish the underestimation of variability in the data, *they do not eliminate it.*
- Therefore, inferences from singly imputed data sets are usually liberal:
  - SEs are underestimated,
  - test statistics are too large,
  - $p$-values are too small,
  - confidence intervals are too tight.

A solution to the underestimation of variability associated with single imputation is to do it repeatedly, capturing the variability from one imputation to another. Such an approach is called *multiple imputation* or MI.

- In R, the most popular package for MI is `mice`.

# Multiple Imputation

- Although stochastic imputation methods diminish the underestimation of variability in the data, *they do not eliminate it.*
- Therefore, inferences from singly imputed data sets are usually liberal:
  - SEs are underestimated,
  - test statistics are too large,
  - $p$-values are too small,
  - confidence intervals are too tight.

A solution to the underestimation of variability associated with single imputation is to do it repeatedly, capturing the variability from one imputation to another. Such an approach is called *multiple imputation* or MI.

- In R, the most popular package for MI is `mice`.

# Multiple Imputation

- Although stochastic imputation methods diminish the underestimation of variability in the data, *they do not eliminate it.*
- Therefore, inferences from singly imputed data sets are usually liberal:
  - SEs are underestimated,
  - test statistics are too large,
  - $p$-values are too small,
  - confidence intervals are too tight.

A solution to the underestimation of variability associated with single imputation is to do it repeatedly, capturing the variability from one imputation to another. Such an approach is called *multiple imputation* or MI.

- In R, the most popular package for MI is `mice`.

# Multiple Imputation

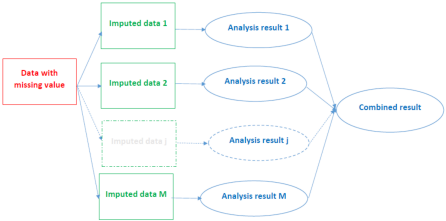



Figure 1: Figure from the UCLA IDRE seminar, Missing Data in R.

MI is a three-step process:

1. **Multiply Impute:** Stochastic imputation is used to create $M$ complete data sets.
2. **Analyze:** Analyze each data set as you would if there had been no missing data.
3. **Pool:** Results are pooled to create a single result.

# Multiple Imputation

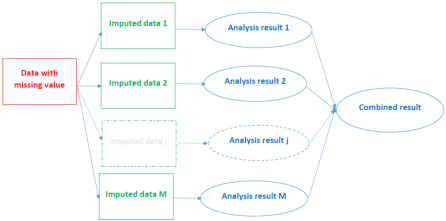



Figure 1: Figure from the UCLA IDRE seminar, Missing Data in R.

MI is a three-step process:

1. **Multiply Impute:** Stochastic imputation is used to create $M$ complete data sets.
2. **Analyze:** Analyze each data set as you would if there had been no missing data.
3. **Pool:** Results are pooled to create a single result.

# Multiple Imputation

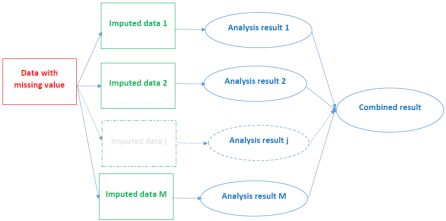



Figure 1: Figure from the UCLA IDRE seminar, Missing Data in R.

MI is a three-step process:

1. **Multiply Impute:** Stochastic imputation is used to create $M$ complete data sets.
2. **Analyze:** Analyze each data set as you would if there had been no missing data.
3. **Pool:** Results are pooled to create a single result.

# Multiple Imputation

When multiple variables have missingness, imputation can be done in three ways:

1. Joint Modeling. Joint model assumed for all variables involved in imputation. Only feasible model is multivariate normal, which is not ideal for discrete variables.
2. Monotone Imputation. only applies if missingness pattern is monotone. In this case, imputation can be done in one pass through the variables.
   - Flexible and quick. But more potential for biased imputations.
3. Chained Equations (MICE algorithm). Each variable's missingness imputed based on sequential regressions involving (usually) all other variables, including those with missing values. Initial imputations are needed and then improved by fitting these sequential regressions iteratively.
   - Flexible, efficient, less potential for bias. Default in `mice`.

# Multiple Imputation

When multiple variables have missingness, imputation can be done in three ways:

1. Joint Modeling. Joint model assumed for all variables involved in imputation. Only feasible model is multivariate normal, which is not ideal for discrete variables.
2. Monotone Imputation. only applies if missingness pattern is monotone. In this case, imputation can be done in one pass through the variables.
   - Flexible and quick. But more potential for biased imputations.
3. Chained Equations (MICE algorithm). Each variable's missingness imputed based on sequential regressions involving (usually) all other variables, including those with missing values. Initial imputations are needed and then improved by fitting these sequential regressions iteratively.
   - Flexible, efficient, less potential for bias. Default in `mice`.

# Multiple Imputation

When multiple variables have missingness, imputation can be done in three ways:

1. Joint Modeling. Joint model assumed for all variables involved in imputation. Only feasible model is multivariate normal, which is not ideal for discrete variables.
2. Monotone Imputation. only applies if missingness pattern is monotone. In this case, imputation can be done in one pass through the variables.
   - Flexible and quick. But more potential for biased imputations.
3. Chained Equations (MICE algorithm). Each variable's missingness imputed based on sequential regressions involving (usually) all other variables, including those with missing values. Initial imputations are needed and then improved by fitting these sequential regressions iteratively.
   - Flexible, efficient, less potential for bias. Default in `mice`.

# Multiple Imputation

When multiple variables have missingness, imputation can be done in three ways:

1. Joint Modeling. Joint model assumed for all variables involved in imputation. Only feasible model is multivariate normal, which is not ideal for discrete variables.
2. Monotone Imputation. only applies if missingness pattern is monotone. In this case, imputation can be done in one pass through the variables.
   - Flexible and quick. But more potential for biased imputations.
3. Chained Equations (MICE algorithm). Each variable's missingness imputed based on sequential regressions involving (usually) all other variables, including those with missing values. Initial imputations are needed and then improved by fitting these sequential regressions iteratively.
   - Flexible, efficient, less potential for bias. Default in `mice`.

# Multiple Imputation

When multiple variables have missingness, imputation can be done in three ways:

1. Joint Modeling. Joint model assumed for all variables involved in imputation. Only feasible model is multivariate normal, which is not ideal for discrete variables.
2. Monotone Imputation. only applies if missingness pattern is monotone. In this case, imputation can be done in one pass through the variables.
    - Flexible and quick. But more potential for biased imputations.
3. Chained Equations (MICE algorithm). Each variable's missingness imputed based on sequential regressions involving (usually) all other variables, including those with missing values. Initial imputations are needed and then improved by fitting these sequential regressions iteratively.
    - Flexible, efficient, less potential for bias. Default in `mice`.

# Full Information Maximum Likelihood Estimation (FIML)

- Maximum likelihood estimation is a widely used method of estimation for statistical models with an associated suite of large-sample inference methods.
- When applied to fit models to available cases when missingness is present, it is often called FIML estimation.
- FIML requires a joint model for all of the variables of interest under complete data. This model describes the distribution of the response variable(s) and, if covariates have missing data, the covariates.
- Under such a model, it is possible to write down the joint probability density (also known as the likelihood function) of the observed incomplete data.
- In general, this requires summing (for discrete data) or integrating (for continuous data) the missing variables out of the complete data joint density function.
- If we can evaluate the likelihood of the observed data, then we can numerically maximize that likelihood to do parameter estimation.

# Full Information Maximum Likelihood Estimation (FIML)

- Maximum likelihood estimation is a widely used method of estimation for statistical models with an associated suite of large-sample inference methods.
- When applied to fit models to available cases when missingness is present, it is often called FIML estimation.
- FIML requires a joint model for all of the variables of interest under complete data. This model describes the distribution of the response variable(s) and, if covariates have missing data, the covariates.
- Under such a model, it is possible to write down the joint probability density (also known as the likelihood function) of the observed incomplete data.
- In general, this requires summing (for discrete data) or integrating (for continuous data) the missing variables out of the complete data joint density function.
- If we can evaluate the likelihood of the observed data, then we can numerically maximize that likelihood to do parameter estimation.

# Full Information Maximum Likelihood Estimation (FIML)

- Maximum likelihood estimation is a widely used method of estimation for statistical models with an associated suite of large-sample inference methods.
- When applied to fit models to available cases when missingness is present, it is often called FIML estimation.
- FIML requires a joint model for all of the variables of interest under complete data. This model describes the distribution of the response variable(s) and, if covariates have missing data, the covariates.
- Under such a model, it is possible to write down the joint probability density (also known as the likelihood function) of the observed incomplete data.
- In general, this requires summing (for discrete data) or integrating (for continuous data) the missing variables out of the complete data joint density function.
- If we can evaluate the likelihood of the observed data, then we can numerically maximize that likelihood to do parameter estimation.

# Full Information Maximum Likelihood Estimation (FIML)

- Maximum likelihood estimation is a widely used method of estimation for statistical models with an associated suite of large-sample inference methods.
- When applied to fit models to available cases when missingness is present, it is often called FIML estimation.
- FIML requires a joint model for all of the variables of interest under complete data. This model describes the distribution of the response variable(s) and, if covariates have missing data, the covariates.
- Under such a model, it is possible to write down the joint probability density (also known as the likelihood function) of the observed incomplete data.
- In general, this requires summing (for discrete data) or integrating (for continuous data) the missing variables out of the complete data joint density function.
- If we can evaluate the likelihood of the observed data, then we can numerically maximize that likelihood to do parameter estimation.

# Full Information Maximum Likelihood Estimation (FIML)

- Maximum likelihood estimation is a widely used method of estimation for statistical models with an associated suite of large-sample inference methods.
- When applied to fit models to available cases when missingness is present, it is often called FIML estimation.
- FIML requires a joint model for all of the variables of interest under complete data. This model describes the distribution of the response variable(s) and, if covariates have missing data, the covariates.
- Under such a model, it is possible to write down the joint probability density (also known as the likelihood function) of the observed incomplete data.
- In general, this requires summing (for discrete data) or integrating (for continuous data) the missing variables out of the complete data joint density function.
- If we can evaluate the likelihood of the observed data, then we can numerically maximize that likelihood to do parameter estimation.

# Full Information Maximum Likelihood Estimation (FIML)

- Maximum likelihood estimation is a widely used method of estimation for statistical models with an associated suite of large-sample inference methods.
- When applied to fit models to available cases when missingness is present, it is often called FIML estimation.
- FIML requires a joint model for all of the variables of interest under complete data. This model describes the distribution of the response variable(s) and, if covariates have missing data, the covariates.
- Under such a model, it is possible to write down the joint probability density (also known as the likelihood function) of the observed incomplete data.
- In general, this requires summing (for discrete data) or integrating (for continuous data) the missing variables out of the complete data joint density function.
- If we can evaluate the likelihood of the observed data, then we can numerically maximize that likelihood to do parameter estimation.

# Full Information Maximum Likelihood Estimation (FIML)

- If the missingness is restricted to the response variable(s), no "extra" model is needed and FIML is straight-forward without specialized software.
- However, if predictors have missing values, then a model for the joint distribution of those predictor and the response(s) is needed, which would not be required for a standard regression analysis, say.
- Moreover, specialized software is typically needed. Most software for structural equation models and other latent variable models.
  - The `lavaan` package in R implements FIML for many models.

# Full Information Maximum Likelihood Estimation (FIML)

- If the missingness is restricted to the response variable(s), no "extra" model is needed and FIML is straight-forward without specialized software.
- However, if predictors have missing values, then a model for the joint distribution of those predictor and the response(s) is needed, which would not be required for a standard regression analysis, say.
- Moreover, specialized software is typically needed. Most software for structural equation models and other latent variable models.
  - The `lavaan` package in R implements FIML for many models.

# Full Information Maximum Likelihood Estimation (FIML)

- If the missingness is restricted to the response variable(s), no "extra" model is needed and FIML is straight-forward without specialized software.
- However, if predictors have missing values, then a model for the joint distribution of those predictor and the response(s) is needed, which would not be required for a standard regression analysis, say.
- Moreover, specialized software is typically needed. Most software for structural equation models and other latent variable models.
  - The `lavaan` package in R implements FIML for many models.

# Full Information Maximum Likelihood Estimation (FIML)

- If the missingness is restricted to the response variable(s), no "extra" model is needed and FIML is straight-forward without specialized software.
- However, if predictors have missing values, then a model for the joint distribution of those predictor and the response(s) is needed, which would not be required for a standard regression analysis, say.
- Moreover, specialized software is typically needed. Most software for structural equation models and other latent variable models.
  - The `lavaan` package in R implements FIML for many models.

# Recommendations

- In regression, include covariates that may be related to missingness even if not related to the response. This makes MAR more plausible.

1. MCAR:
   - Do not assume without diagnostics.
   - Complete case analysis ok, but inefficient. If data set is large, inefficiency is not of concern. Otherwise, consider MI or FIML.
2. MAR:
   - If missingness only in $Y$, complete case analysis is sufficient.
   - Otherwise, if covariates have missing data, MI or FIML estimation may be a good choice. But throw out covariates with a high proportion of missing values.
3. NMAR
   - Must model the missing data mechanism and link that model to the model for the response.
   - Problem: NMAR cannot be diagnosed from observed data. So, model choice is speculative.

# Recommendations

- In regression, include covariates that may be related to missingness even if not related to the response. This makes MAR more plausible.

1. MCAR:
   - Do not assume without diagnostics.
   - Complete case analysis ok, but inefficient. If data set is large, inefficiency is not of concern. Otherwise, consider MI or FIML.
2. MAR:
   - If missingness only in $Y$, complete case analysis is sufficient.
   - Otherwise, if covariates have missing data, MI or FIML estimation may be a good choice. But throw out covariates with a high proportion of missing values.
   - Note MAR is right? I suppose, 99% but never completely sure.
3. NMAR
   - Must model the missing data mechanism and link that model to the model for the response.
   - Problem: NMAR cannot be diagnosed from observed data. So, model choice is speculative.

# Recommendations

- In regression, include covariates that may be related to missingness even if not related to the response. This makes MAR more plausible.

1. MCAR:
   - Do not assume without diagnostics.
   - Complete case analysis ok, but inefficient. If data set is large, inefficiency is not of concern. Otherwise, consider MI or FIML.
2. MAR:
   - If missingness only in $Y$, complete case analysis is sufficient.
   - Otherwise, if covariates have missing data, MI or FIML estimation may be a good choice. But throw out covariates with a high proportion of missing values.
   - Note MAR is right? I guess, 50% but it is complete bollox.
3. NMAR
   - Must model the missing data mechanism and link that model to the model for the response.
   - Problem: NMAR cannot be diagnosed from observed data. So, model choice is speculative.

# Recommendations

- In regression, include covariates that may be related to missingness even if not related to the response. This makes MAR more plausible.

1. MCAR:
   - Do not assume without diagnostics.
   - Complete case analysis ok, but inefficient. If data set is large, inefficiency is not of concern. Otherwise, consider MI or FIML.
2. MAR:
   - If missingness only in $Y$, complete case analysis is sufficient.
   - Otherwise, if covariates have missing data, MI or FIML estimation may be a good choice. But throw out covariates with a high proportion of missing values.
   - Note MAR is right? I guess, 50% but it's complicated. Rubin
3. NMAR
   - Must model the missing data mechanism and link that model to the model for the response.
   - Problem: NMAR cannot be diagnosed from observed data. So, model choice is speculative.

# Recommendations

- In regression, include covariates that may be related to missingness even if not related to the response. This makes MAR more plausible.

1. MCAR:
   - Do not assume without diagnostics.
   - Complete case analysis ok, but inefficient. If data set is large, inefficiency is not of concern. Otherwise, consider MI or FIML.
2. MAR:
   - If missingness only in $Y$, complete case analysis is sufficient.
   - Otherwise, if covariates have missing data, MI or FIML estimation may be a good choice. But throw out covariates with a high proportion of missing values.
     - How high is high? I suggest 50%, but more comments below.
3. NMAR
   - Must model the missing data mechanism and link that model to the model for the response.
   - Problem: NMAR cannot be diagnosed from observed data. So, model choice is speculative.

# Recommendations

- In regression, include covariates that may be related to missingness even if not related to the response. This makes MAR more plausible.

1. MCAR:
   - Do not assume without diagnostics.
   - Complete case analysis ok, but inefficient. If data set is large, inefficiency is not of concern. Otherwise, consider MI or FIML.
2. MAR:
   - If missingness only in $Y$, complete case analysis is sufficient.
   - Otherwise, if covariates have missing data, MI or FIML estimation may be a good choice. But throw out covariates with a high proportion of missing values.
     - How high is high? I suggest 50%, but more comments below.
3. NMAR
   - Must model the missing data mechanism and link that model to the model for the response.
   - Problem: NMAR cannot be diagnosed from observed data. So, model choice is speculative.

# Recommendations

- In regression, include covariates that may be related to missingness even if not related to the response. This makes MAR more plausible.

1. MCAR:
   - Do not assume without diagnostics.
   - Complete case analysis ok, but inefficient. If data set is large, inefficiency is not of concern. Otherwise, consider MI or FIML.
2. MAR:
   - If missingness only in $Y$, complete case analysis is sufficient.
   - Otherwise, if covariates have missing data, MI or FIML estimation may be a good choice. But throw out covariates with a high proportion of missing values.
     - How high is high? I suggest 50%, but more comments below.
3. NMAR
   - Must model the missing data mechanism and link that model to the model for the response.
   - Problem: NMAR cannot be diagnosed from observed data. So, model choice is speculative.

# Recommendations

- In regression, include covariates that may be related to missingness even if not related to the response. This makes MAR more plausible.

1. MCAR:
   - Do not assume without diagnostics.
   - Complete case analysis ok, but inefficient. If data set is large, inefficiency is not of concern. Otherwise, consider MI or FIML.
2. MAR:
   - If missingness only in $Y$, complete case analysis is sufficient.
   - Otherwise, if covariates have missing data, MI or FIML estimation may be a good choice. But throw out covariates with a high proportion of missing values.
     - How high is high? I suggest 50%, but more comments below.
3. NMAR
   - Must model the missing data mechanism and link that model to the model for the response.
   - Problem: NMAR cannot be diagnosed from observed data. So, model choice is speculative.

# Recommendations

- In regression, include covariates that may be related to missingness even if not related to the response. This makes MAR more plausible.

1. MCAR:
   - Do not assume without diagnostics.
   - Complete case analysis ok, but inefficient. If data set is large, inefficiency is not of concern. Otherwise, consider MI or FIML.
2. MAR:
   - If missingness only in $Y$, complete case analysis is sufficient.
   - Otherwise, if covariates have missing data, MI or FIML estimation may be a good choice. But throw out covariates with a high proportion of missing values.
     - How high is high? I suggest 50%, but more comments below.
3. NMAR
   - Must model the missing data mechanism and link that model to the model for the response.
   - Problem: NMAR cannot be diagnosed from observed data. So, model choice is speculative.

# Recommendations

- In regression, include covariates that may be related to missingness even if not related to the response. This makes MAR more plausible.

1. MCAR:
   - Do not assume without diagnostics.
   - Complete case analysis ok, but inefficient. If data set is large, inefficiency is not of concern. Otherwise, consider MI or FIML.
2. MAR:
   - If missingness only in $Y$, complete case analysis is sufficient.
   - Otherwise, if covariates have missing data, MI or FIML estimation may be a good choice. But throw out covariates with a high proportion of missing values.
     - How high is high? I suggest 50%, but more comments below.
3. NMAR
   - Must model the missing data mechanism and link that model to the model for the response.
   - Problem: NMAR cannot be diagnosed from observed data. So, model choice is speculative.

# Recommendations

- In regression, include covariates that may be related to missingness even if not related to the response. This makes MAR more plausible.

1. MCAR:
   - Do not assume without diagnostics.
   - Complete case analysis ok, but inefficient. If data set is large, inefficiency is not of concern. Otherwise, consider MI or FIML.
2. MAR:
   - If missingness only in $Y$, complete case analysis is sufficient.
   - Otherwise, if covariates have missing data, MI or FIML estimation may be a good choice. But throw out covariates with a high proportion of missing values.
     - How high is high? I suggest 50%, but more comments below.
3. NMAR
   - Must model the missing data mechanism and link that model to the model for the response.
   - Problem: NMAR cannot be diagnosed from observed data. So, model choice is speculative.

# Recommendations

How much missingness is too much?

- If assumptions of FIML or MI are poor, these methods can do more harm then good, especially when missingness is high.
- FIML and MI based on a joint model both assume multivariate normality. For discrete variables, this sometimes works ok, but not if proportion of missingness is high.
- In MICE, assumptions are mild and flexible. Can be confident in this method when missingness is high and/or with mixed-scale variables.
- Effects of covariates with high missingness will be estimated with much uncertainty and are unlikely to be significant in the analysis model. But their presence may improve the imputation or inference on other covariates. So, some experts recommend retaining covariates with >50% missing data.
- If a covariate is of central interest, tolerate more missingness in it.
- Throw away redundant covariates with missingness.

# Recommendations

How much missingness is too much?

- If assumptions of FIML or MI are poor, these methods can do more harm then good, especially when missingness is high.
- FIML and MI based on a joint model both assume multivariate normality. For discrete variables, this sometimes works ok, but not if proportion of missingness is high.
- In MICE, assumptions are mild and flexible. Can be confident in this method when missingness is high and/or with mixed-scale variables.
- Effects of covariates with high missingness will be estimated with much uncertainty and are unlikely to be significant in the analysis model. But their presence may improve the imputation or inference on other covariates. So, some experts recommend retaining covariates with >50% missing data.
- If a covariate is of central interest, tolerate more missingness in it.
- Throw away redundant covariates with missingness.

# Recommendations

How much missingness is too much?

- If assumptions of FIML or MI are poor, these methods can do more harm then good, especially when missingness is high.
- FIML and MI based on a joint model both assume multivariate normality. For discrete variables, this sometimes works ok, but not if proportion of missingness is high.
- In MICE, assumptions are mild and flexible. Can be confident in this method when missingness is high and/or with mixed-scale variables.
- Effects of covariates with high missingness will be estimated with much uncertainty and are unlikely to be significant in the analysis model. But their presence may improve the imputation or inference on other covariates. So, some experts recommend retaining covariates with >50% missing data.
- If a covariate is of central interest, tolerate more missingness in it.
- Throw away redundant covariates with missingness.

# Recommendations

How much missingness is too much?

- If assumptions of FIML or MI are poor, these methods can do more harm then good, especially when missingness is high.
- FIML and MI based on a joint model both assume multivariate normality. For discrete variables, this sometimes works ok, but not if proportion of missingness is high.
- In MICE, assumptions are mild and flexible. Can be confident in this method when missingness is high and/or with mixed-scale variables.
- Effects of covariates with high missingness will be estimated with much uncertainty and are unlikely to be significant in the analysis model. But their presence may improve the imputation or inference on other covariates. So, some experts recommend retaining covariates with >50% missing data.
- If a covariate is of central interest, tolerate more missingness in it.
- Throw away redundant covariates with missingness.

# Recommendations

How much missingness is too much?

- If assumptions of FIML or MI are poor, these methods can do more harm then good, especially when missingness is high.
- FIML and MI based on a joint model both assume multivariate normality. For discrete variables, this sometimes works ok, but not if proportion of missingness is high.
- In MICE, assumptions are mild and flexible. Can be confident in this method when missingness is high and/or with mixed-scale variables.
- Effects of covariates with high missingness will be estimated with much uncertainty and are unlikely to be significant in the analysis model. But their presence may improve the imputation or inference on other covariates. So, some experts recommend retaining covariates with >50% missing data.
- If a covariate is of central interest, tolerate more missingness in it.
- Throw away redundant covariates with missingness.

# Recommendations

How much missingness is too much?

- If assumptions of FIML or MI are poor, these methods can do more harm then good, especially when missingness is high.
- FIML and MI based on a joint model both assume multivariate normality. For discrete variables, this sometimes works ok, but not if proportion of missingness is high.
- In MICE, assumptions are mild and flexible. Can be confident in this method when missingness is high and/or with mixed-scale variables.
- Effects of covariates with high missingness will be estimated with much uncertainty and are unlikely to be significant in the analysis model. But their presence may improve the imputation or inference on other covariates. So, some experts recommend retaining covariates with >50% missing data.
- If a covariate is of central interest, tolerate more missingness in it.
- Throw away redundant covariates with missingness.

# Final Comments

MI is best suited to fitting and doing inference on a single model. For complex analyses, many questions arise.

- What if we need to do variable selection?
- How do we impute missing values for structured data (time series, longitudinal/multilevel data, Spatio-temporal data)?
- How do we handle non-convergence, quasi-complete separation, multicollinearity in imputation models or analysis models for some imputed data sets?
- How do we pool exact $p$-values or CIs (as obtained in exact conditional logistic regression, say).
- How do we do analysis model diagnostics? Is it sufficient to do them on one imputed data set?

FIML is simpler than MI in many respects and avoids many of its complications. But it still poses practical difficulties. Two main concerns:

- it is based on stronger assumptions that are not suitable for many problems, and
- implementation is limited and/or unfamiliar.

# Final Comments

MI is best suited to fitting and doing inference on a single model. For complex analyses, many questions arise.

- What if we need to do variable selection?
- How do we impute missing values for structured data (time series, longitudinal/multilevel data, Spatio-temporal data)?
- How do we handle non-convergence, quasi-complete separation, multicollinearity in imputation models or analysis models for some imputed data sets?
- How do we pool exact $p$-values or CIs (as obtained in exact conditional logistic regression, say).
- How do we do analysis model diagnostics? Is it sufficient to do them on one imputed data set?

FIML is simpler than MI in many respects and avoids many of its complications. But it still poses practical difficulties. Two main concerns:

- it is based on stronger assumptions that are not suitable for many problems, and
- implementation is limited and/or unfamiliar.

# Final Comments

MI is best suited to fitting and doing inference on a single model. For complex analyses, many questions arise.

- What if we need to do variable selection?
- How do we impute missing values for structured data (time series, longitudinal/multilevel data, Spatio-temporal data)?
- How do we handle non-convergence, quasi-complete separation, multicollinearity in imputation models or analysis models for some imputed data sets?
- How do we pool exact $p$-values or CIs (as obtained in exact conditional logistic regression, say).
- How do we do analysis model diagnostics? Is it sufficient to do them on one imputed data set?

FIML is simpler than MI in many respects and avoids many of its complications. But it still poses practical difficulties. Two main concerns:

- it is based on stronger assumptions that are not suitable for many problems, and
- implementation is limited and/or unfamiliar.

# Final Comments

MI is best suited to fitting and doing inference on a single model. For complex analyses, many questions arise.

- What if we need to do variable selection?
- How do we impute missing values for structured data (time series, longitudinal/multilevel data, Spatio-temporal data)?
- How do we handle non-convergence, quasi-complete separation, multicollinearity in imputation models or analysis models for some imputed data sets?
- How do we pool exact $p$-values or CIs (as obtained in exact conditional logistic regression, say).
- How do we do analysis model diagnostics? Is it sufficient to do them on one imputed data set?

FIML is simpler than MI in many respects and avoids many of its complications. But it still poses practical difficulties. Two main concerns:

- it is based on stronger assumptions that are not suitable for many problems, and
- implementation is limited and/or unfamiliar.

# Final Comments

MI is best suited to fitting and doing inference on a single model. For complex analyses, many questions arise.

- What if we need to do variable selection?
- How do we impute missing values for structured data (time series, longitudinal/multilevel data, Spatio-temporal data)?
- How do we handle non-convergence, quasi-complete separation, multicollinearity in imputation models or analysis models for some imputed data sets?
- How do we pool exact $p$-values or CIs (as obtained in exact conditional logistic regression, say).
- How do we do analysis model diagnostics? Is it sufficient to do them on one imputed data set?

FIML is simpler than MI in many respects and avoids many of its complications. But it still poses practical difficulties. Two main concerns:

- it is based on stronger assumptions that are not suitable for many problems, and
- implementation is limited and/or unfamiliar.

# Final Comments

MI is best suited to fitting and doing inference on a single model. For complex analyses, many questions arise.

- What if we need to do variable selection?
- How do we impute missing values for structured data (time series, longitudinal/multilevel data, Spatio-temporal data)?
- How do we handle non-convergence, quasi-complete separation, multicollinearity in imputation models or analysis models for some imputed data sets?
- How do we pool exact $p$-values or CIs (as obtained in exact conditional logistic regression, say).
- How do we do analysis model diagnostics? Is it sufficient to do them on one imputed data set?

FIML is simpler than MI in many respects and avoids many of its complications. But it still poses practical difficulties. Two main concerns:

- it is based on stronger assumptions that are not suitable for many problems, and
- implementation is limited and/or unfamiliar.

# Final Comments

MI is best suited to fitting and doing inference on a single model. For complex analyses, many questions arise.

- What if we need to do variable selection?
- How do we impute missing values for structured data (time series, longitudinal/multilevel data, Spatio-temporal data)?
- How do we handle non-convergence, quasi-complete separation, multicollinearity in imputation models or analysis models for some imputed data sets?
- How do we pool exact $p$-values or CIs (as obtained in exact conditional logistic regression, say).
- How do we do analysis model diagnostics? Is it sufficient to do them on one imputed data set?

FIML is simpler than MI in many respects and avoids many of its complications. But it still poses practical difficulties. Two main concerns:

- it is based on stronger assumptions that are not suitable for many problems, and
- implementation is limited and/or unfamiliar.

# References & Resources

See `MissingDataNotes.html` for references, online resources, and a list of usual R packages,

# Thanks

If you need assistance when analyzing data with missingness, please contact the SCC.

- www.stat.uga/consulting

We can help!

Thank you!

# Questions?

Questions?