Missing Data

Rachel Pompa Andy Retting

University of Wyoming

December 12, 2019

Introduction

Introduction

Missing data:

- What it is
- Why it matters
- Examples

Introduction

Missing data:

- What it is
  - Occurs when there is no data value for a variable in an observation
  - Gaps in explanatory variables
  - Gaps in outcome variable
- Why it matters
- Examples

Introduction

Missing data:

- What it is
- Why it matters
  - Occurs in every field that runs empirical research
  - Have to decide how to treat missing data
  - Important to understand missingness for proper handling and analysis
- Examples

Introduction

Missing data:

- What it is
- Why it matters
- Examples
  - Survey Data: Respondents may not answer all the questions due to sensitivity.
  - Time Series: The data doesn't exist in certain frequencies.
  - Panel Data: Gaps arise from attrition in surveys/studies.

Introduction

Why do we care?

Introduction

Why do we care?

- Reduces sample size

Introduction

Why do we care?

- Reduces sample size
  - Widens confidence intervals

Introduction

Why do we care?

- Reduces sample size
  - Widens confidence intervals
  - Reduces statistical power

Introduction

Why do we care?

- Reduces sample size
  - Widens confidence intervals
  - Reduces statistical power
  - Estimators may be biased

Introduction

Why do we care?

- Reduces sample size
  - Widens confidence intervals
  - Reduces statistical power
  - Estimators may be biased
- Limits the ability to observe patterns over time

Introduction

Why do we care?

- Reduces sample size
  - Widens confidence intervals
  - Reduces statistical power
  - Estimators may be biased
- Limits the ability to observe patterns over time
- Harder to draw valid inferences

Outline

- Classifications of missing data
  - Missing completely at Random (MCAR)
  - Missing at Random (MAR)
  - Not Missing at Random (NMAR)
- Methods for dealing with missing data
  - Listwise Deletion
  - Zero-Order Method
  - Modified Zero-Order Regression
- Matlab Code

Classifications of Missing Data

Missing Completely at Random (MCAR):

Classifications of Missing Data

Missing Completely at Random (MCAR):

- Missingness is unrelated to the values of the other observations

Classifications of Missing Data

Missing Completely at Random (MCAR):

- Missingness is unrelated to the values of the other observations
  - Not systematic

Classifications of Missing Data

Missing Completely at Random (MCAR):

- Missingness is unrelated to the values of the other observations
  - Not systematic
- Usable dataset

Classifications of Missing Data

Missing Completely at Random (MCAR):

- Missingness is unrelated to the values of the other observations
  - Not systematic
- Usable dataset
  - “Ignorable Case”

Classifications of Missing Data

Missing Completely at Random (MCAR):

- Missingness is unrelated to the values of the other observations
  - Not systematic
- Usable dataset
  - “Ignorable Case”
- Relative to a complete dataset

Classifications of Missing Data

Missing Completely at Random (MCAR):

- Missingness is unrelated to the values of the other observations
  - Not systematic
- Usable dataset
  - “Ignorable Case”
- Relative to a complete dataset
  - Estimators are unbiased

Classifications of Missing Data

Missing Completely at Random (MCAR):

- Missingness is unrelated to the values of the other observations
  - Not systematic
- Usable dataset
  - “Ignorable Case”
- Relative to a complete dataset
  - Estimators are unbiased
  - Decreased efficiency

Classifications of Missing Data

Missing Completely at Random (MCAR) Examples:

Classifications of Missing Data

Missing Completely at Random (MCAR) Examples:

- You have a large dataset where several observations become corrupted

Classifications of Missing Data

Missing Completely at Random (MCAR) Examples:

- You have a large dataset where several observations become corrupted
- Your cat walks over your keyboard and hits the delete key without you noticing

Classifications of Missing Data

Missing Completely at Random (MCAR) Examples:

- You have a large dataset where several observations become corrupted
- Your cat walks over your keyboard and hits the delete key without you noticing
- You accidentally miss an observation when entering data

Classifications of Missing Data

Missing at Random (MAR):

Classifications of Missing Data

Missing at Random (MAR):

- Missingness is contained in the complete observations

Classifications of Missing Data

Missing at Random (MAR):

- Missingness is contained in the complete observations
  - Systematic

Classifications of Missing Data

Missing at Random (MAR):

- Missingness is contained in the complete observations
  - Systematic
  - “Intermediate Case”

Classifications of Missing Data

Missing at Random (MAR):

- Missingness is contained in the complete observations
  - Systematic
  - “Intermediate Case”
- Relative to a complete dataset

Classifications of Missing Data

Missing at Random (MAR):

- Missingness is contained in the complete observations
  - Systematic
  - “Intermediate Case”

- Relative to a complete dataset
  - Estimators are unbiased

Classifications of Missing Data

Missing at Random (MAR):

- Missingness is contained in the complete observations
  - Systematic
  - “Intermediate Case”

- Relative to a complete dataset
  - Estimators are unbiased
  - Decreased efficiency

Classifications of Missing Data

Missing at Random (MAR) Examples:

Classifications of Missing Data

Missing at Random (MAR) Examples:

- You collect survey responses, calling only at mid-day, losing students and “9 to 5” workers ...

Classifications of Missing Data

Missing at Random (MAR) Examples:

- You collect survey responses, calling only at mid-day, losing students and “9 to 5” workers ...
  but the survey questionnaire inquires about favorite colors.

Classifications of Missing Data

Not Missing at Random (NMAR):

Classifications of Missing Data

Not Missing at Random (NMAR):

- Missingness doesn't satisfy MCAR or MAR

Classifications of Missing Data

Not Missing at Random (NMAR):

- Missingness doesn't satisfy MCAR or MAR
  - Systematic

Classifications of Missing Data

Not Missing at Random (NMAR):

- Missingness doesn't satisfy MCAR or MAR
  - Systematic
  - Related to the phenomenon being modeled

Classifications of Missing Data

Not Missing at Random (NMAR):

- Missingness doesn't satisfy MCAR or MAR
  - Systematic
  - Related to the phenomenon being modeled
  - "Non-ignorable case"

Classifications of Missing Data

Not Missing at Random (NMAR):

- Missingness doesn't satisfy MCAR or MAR
  - Systematic
  - Related to the phenomenon being modeled
  - "Non-ignorable case"

- Relative to a complete dataset

Classifications of Missing Data

Not Missing at Random (NMAR):

- Missingness doesn't satisfy MCAR or MAR
  - Systematic
  - Related to the phenomenon being modeled
  - "Non-ignorable case"

- Relative to a complete dataset
  - Estimators are biased

Classifications of Missing Data

Not Missing at Random (NMAR):

- Missingness doesn't satisfy MCAR or MAR
  - Systematic
  - Related to the phenomenon being modeled
  - "Non-ignorable case"

- Relative to a complete dataset
  - Estimators are biased
  - Decreased efficiency

Classifications of Missing Data

Not Missing at Random (NMAR):

- Missingness doesn't satisfy MCAR or MAR
  - Systematic
  - Related to the phenomenon being modeled
  - "Non-ignorable case"

- Relative to a complete dataset
  - Estimators are biased
  - Decreased efficiency
  - Missing data is correlated with parameters being estimated

Classifications of Missing Data

Not Missing at Random Examples:

Classifications of Missing Data

Not Missing at Random Examples:

- Religious minorities opting out of answering surveys due to intolerance in the area, when the religious make-up of the community is being estimated

Classifications of Missing Data

Not Missing at Random Examples:

- Religious minorities opting out of answering surveys due to intolerance in the area, when the religious make-up of the community is being estimated
- High-school dropouts preferring not to answer questions about education attainment level, when education attainment is being estimated

Methods for dealing with missing data

Listwise Deletion:

Methods for dealing with missing data

Listwise Deletion:

Methods for dealing with missing data

Listwise Deletion:

- Sample model:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$$

Methods for dealing with missing data

Listwise Deletion:

- Sample model:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$$

- Delete observations that have missing data

Methods for dealing with missing data

Listwise Deletion:

- Sample model:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$$

- Delete observations that have missing data
- Simple

Methods for dealing with missing data

Listwise Deletion:

- Sample model:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$$

- Delete observations that have missing data
- Simple
- Reduces statistical power

Methods for dealing with missing data

Listwise Deletion:

- Sample model:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$$

- Delete observations that have missing data
- Simple
- Reduces statistical power
- Unbiased estimators in MCAR and MAR

Methods for dealing with missing data

Listwise Deletion: (for $x_{2,2}$ missing)

Methods for dealing with missing data

Listwise Deletion: (for $x_{2,2}$ missing)

$$\begin{bmatrix} y_1 \\ \cancel{y_2} \\ y_3 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} 1 \\ \cancel{1} \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} x_{2,1} \\ \cancel{\text{(missing)}} \\ x_{2,3} \\ \vdots \\ x_{2,n} \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} x_{3,1} \\ \cancel{x_{3,2}} \\ x_{3,3} \\ \vdots \\ x_{3,n} \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} \epsilon_1 \\ \cancel{\epsilon_2} \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1}$$

$$\begin{bmatrix} y_1 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}_{(n-1)\times 1} \quad \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{(n-1)\times 1} \quad \begin{bmatrix} x_{2,1} \\ x_{2,3} \\ \vdots \\ x_{1,n} \end{bmatrix}_{(n-1)\times 1} \quad \begin{bmatrix} x_{3,1} \\ x_{3,3} \\ \vdots \\ x_{3,n} \end{bmatrix}_{(n-1)\times 1} \quad \begin{bmatrix} \epsilon_1 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}_{(n-1)\times 1}$$

Methods for dealing with missing data

Listwise Deletion: (for $x_{2,2}$ missing)

$$\begin{bmatrix} y_1 \\ \cancel{y_2} \\ y_3 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} 1 \\ \cancel{1} \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} x_{2,1} \\ \cancel{\text{(missing)}} \\ x_{2,3} \\ \vdots \\ x_{2,n} \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} x_{3,1} \\ \cancel{x_{3,2}} \\ x_{3,3} \\ \vdots \\ x_{3,n} \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} \epsilon_1 \\ \cancel{\epsilon_2} \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1}$$

$$\begin{bmatrix} y_1 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}_{(n-1)\times 1} \quad \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{(n-1)\times 1} \quad \begin{bmatrix} x_{2,1} \\ x_{2,3} \\ \vdots \\ x_{1,n} \end{bmatrix}_{(n-1)\times 1} \quad \begin{bmatrix} x_{3,1} \\ x_{3,3} \\ \vdots \\ x_{3,n} \end{bmatrix}_{(n-1)\times 1} \quad \begin{bmatrix} \epsilon_1 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}_{(n-1)\times 1}$$

Methods for dealing with missing data

Zero-Order Method:

Methods for dealing with missing data

Zero-Order Method:

- Sample model:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$$

Methods for dealing with missing data

Zero-Order Method:

- Sample model:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$$

- Mean/Mode Substitution

Methods for dealing with missing data

Zero-Order Method:

- Sample model:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$$

- Mean/Mode Substitution
- Replace missing x with $\overline{x}$ from observed data

Methods for dealing with missing data

Zero-Order Method:

- Sample model:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$$

- Mean/Mode Substitution
- Replace missing x with $\overline{x}$ from observed data
- Use all data

Methods for dealing with missing data

Zero-Order Method:

- Sample model:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$$

- Mean/Mode Substitution
- Replace missing x with $\overline{x}$ from observed data
- Use all data
- Reduces variability

Methods for dealing with missing data

Zero-Order Method:

- Sample model:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$$

- Mean/Mode Substitution
- Replace missing x with $\overline{x}$ from observed data
- Use all data
- Reduces variability
- Weakens covariance and correlation estimates

Methods for dealing with missing data

Zero-Order Method: (for $x_{2,2}$ missing)

Methods for dealing with missing data

Zero-Order Method: (for $x_{2,2}$ missing)

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} x_{2,1} \\ (\text{missing}) \\ x_{2,3} \\ \vdots \\ x_{2,n} \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} x_{3,1} \\ x_{3,2} \\ x_{3,3} \\ \vdots \\ x_{3,n} \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} x_{2,1} \\ \bar{x}_2 \\ x_{2,3} \\ \vdots \\ x_{1,n} \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} x_{3,1} \\ x_{3,2} \\ x_{3,3} \\ \vdots \\ x_{3,n} \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1}$$

Methods for dealing with missing data

Modified Zero-Order Method:

Methods for dealing with missing data

Modified Zero-Order Method:

Methods for dealing with missing data

Modified Zero-Order Method:

- Sample model:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 D_i + \epsilon_i$$

Methods for dealing with missing data

Modified Zero-Order Method:

- Sample model:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 D_i + \epsilon_i$$

- Dummy variable adjustment

Methods for dealing with missing data

Modified Zero-Order Method:

- Sample model:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 D_i + \epsilon_i$$

- Dummy variable adjustment
  - 1 = value is missing observation and 0 = value is not missing

Methods for dealing with missing data

Modified Zero-Order Method:

- Sample model:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 D_i + \epsilon_i$$

- Dummy variable adjustment
  - 1 = value is missing observation and 0 = value is not missing
  - Impute missing values as 0

Methods for dealing with missing data

Modified Zero-Order Method:

- Sample model:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 D_i + \epsilon_i$$

- Dummy variable adjustment
  - 1 = value is missing observation and 0 = value is not missing
  - Impute missing values as 0

- Uses all information

Methods for dealing with missing data

Modified Zero-Order Method: (for $x_{2,2}$ missing)

Methods for dealing with missing data

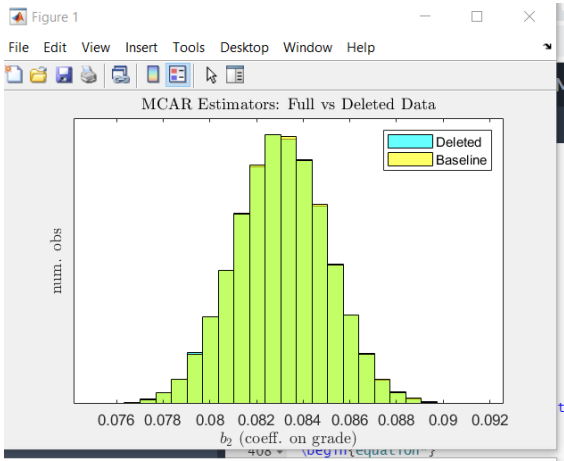
Modified Zero-Order Method: (for $x_{2,2}$ missing)

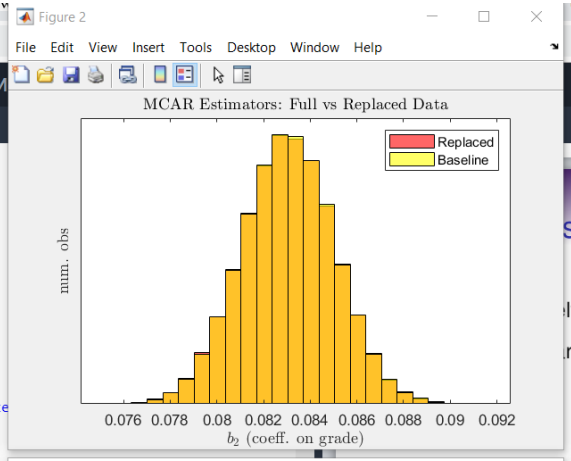
$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} x_{2,1} \\ (\text{missing}) \\ x_{2,3} \\ \vdots \\ x_{2,n} \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} x_{3,1} \\ x_{3,2} \\ x_{3,3} \\ \vdots \\ x_{3,n} \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1}$$

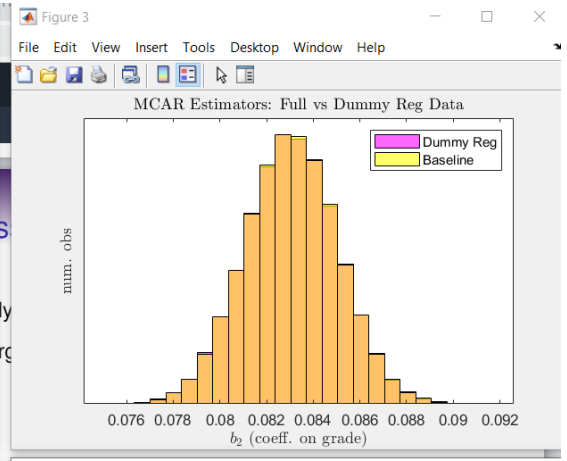
$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} x_{2,1} \\ 0 \\ x_{2,3} \\ \vdots \\ x_{1,n} \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} x_{3,1} \\ x_{3,2} \\ x_{3,3} \\ \vdots \\ x_{3,n} \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{n\times 1} \quad \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1}$$

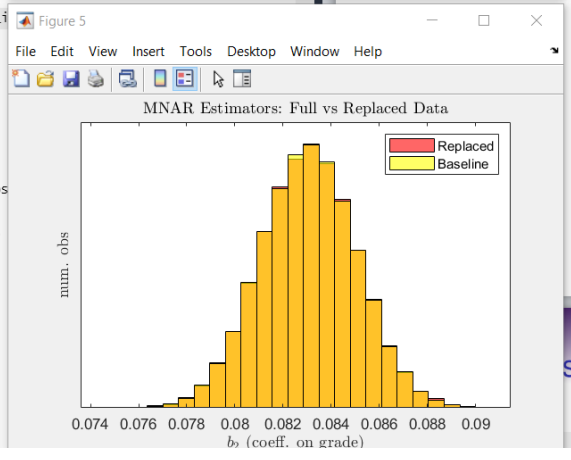
Matlab Graphs

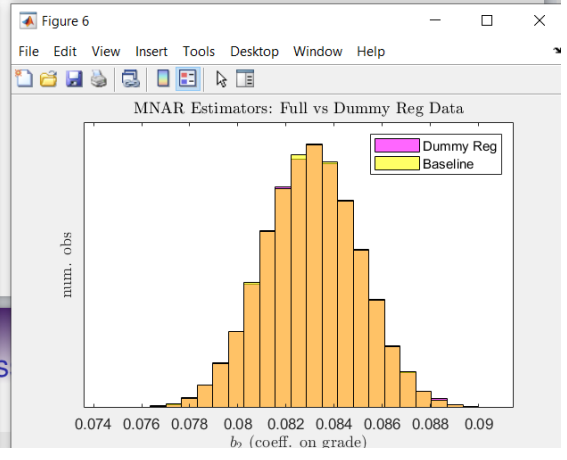
Missing Observations = 1





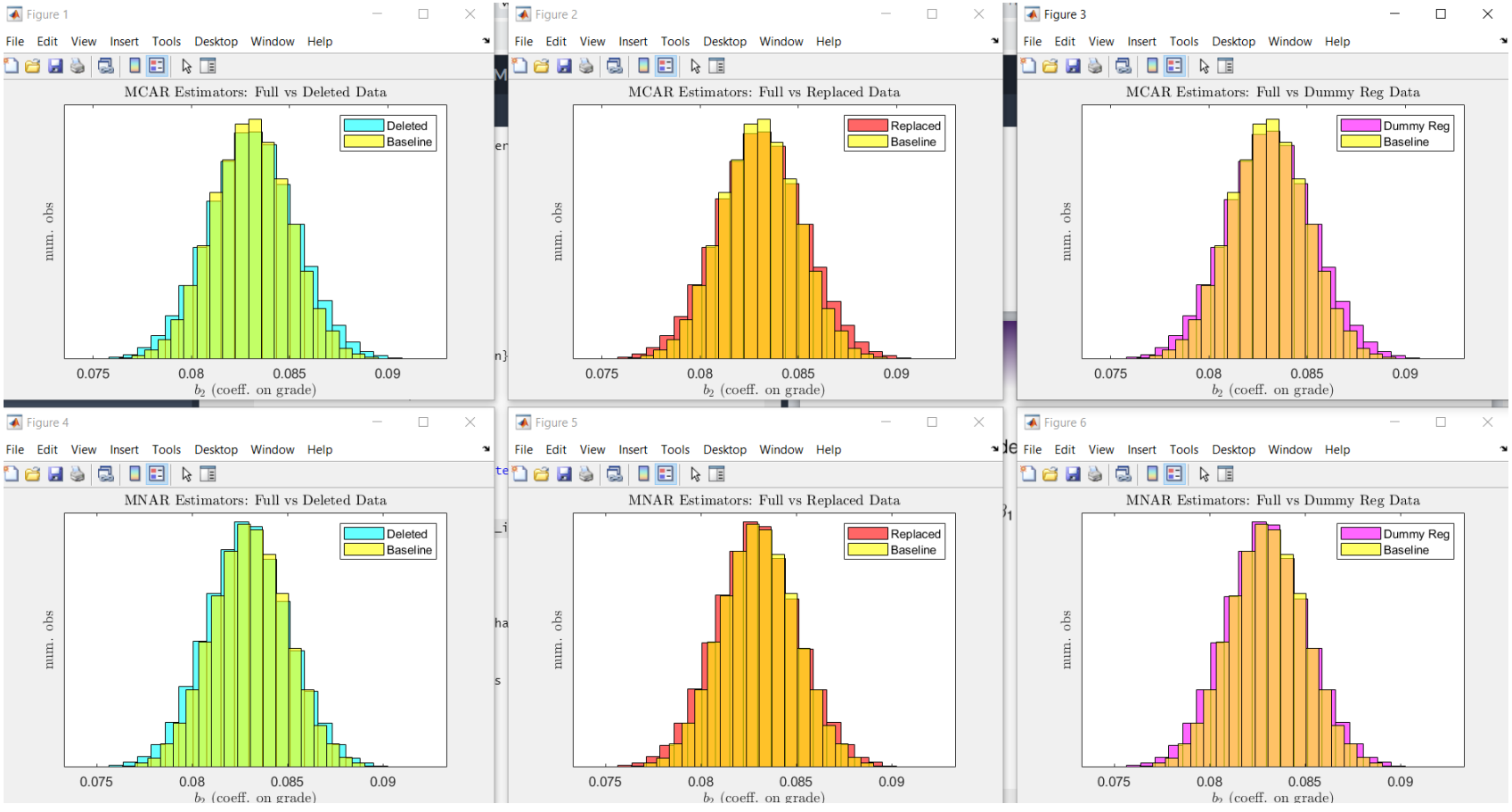






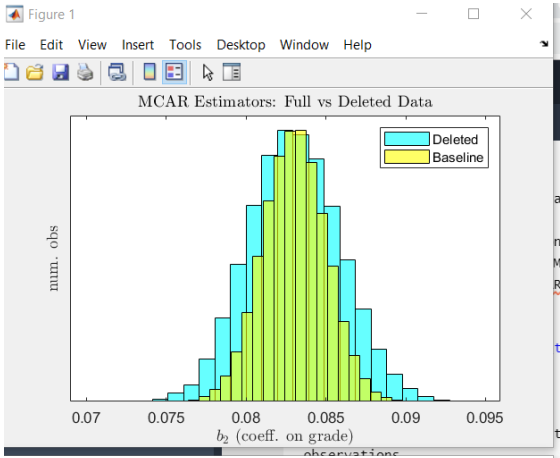
Matlab Graphs

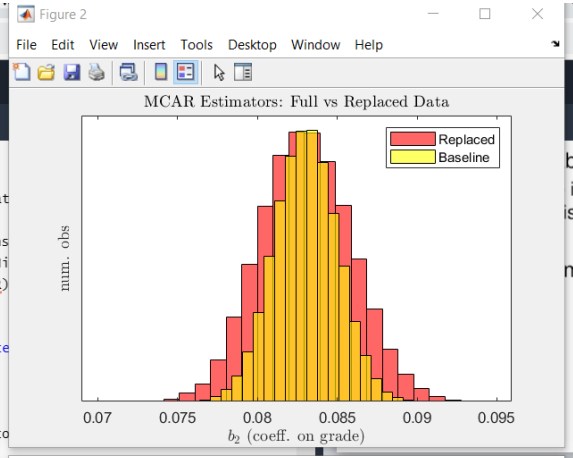
Missing Observations = 25

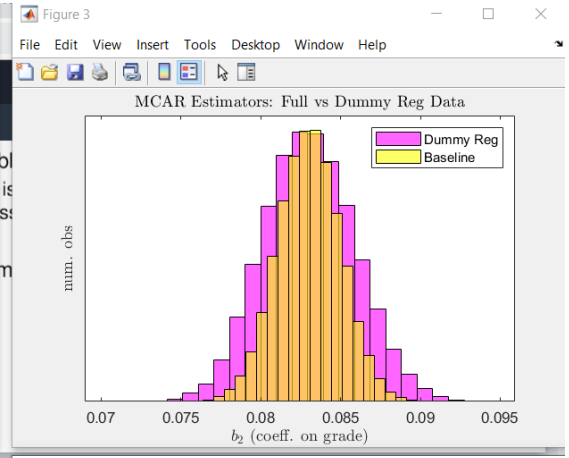


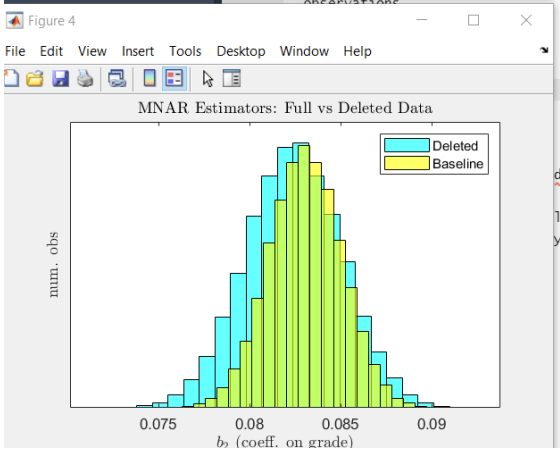
Matlab Graphs

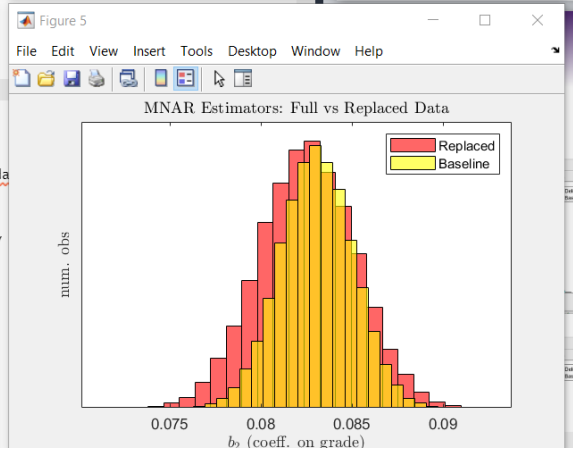
Missing Observations = 100

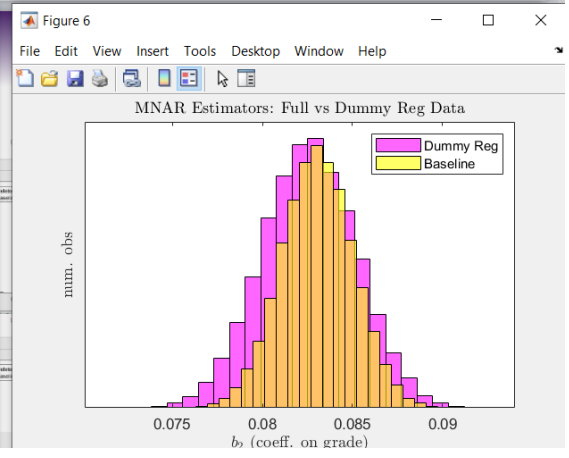






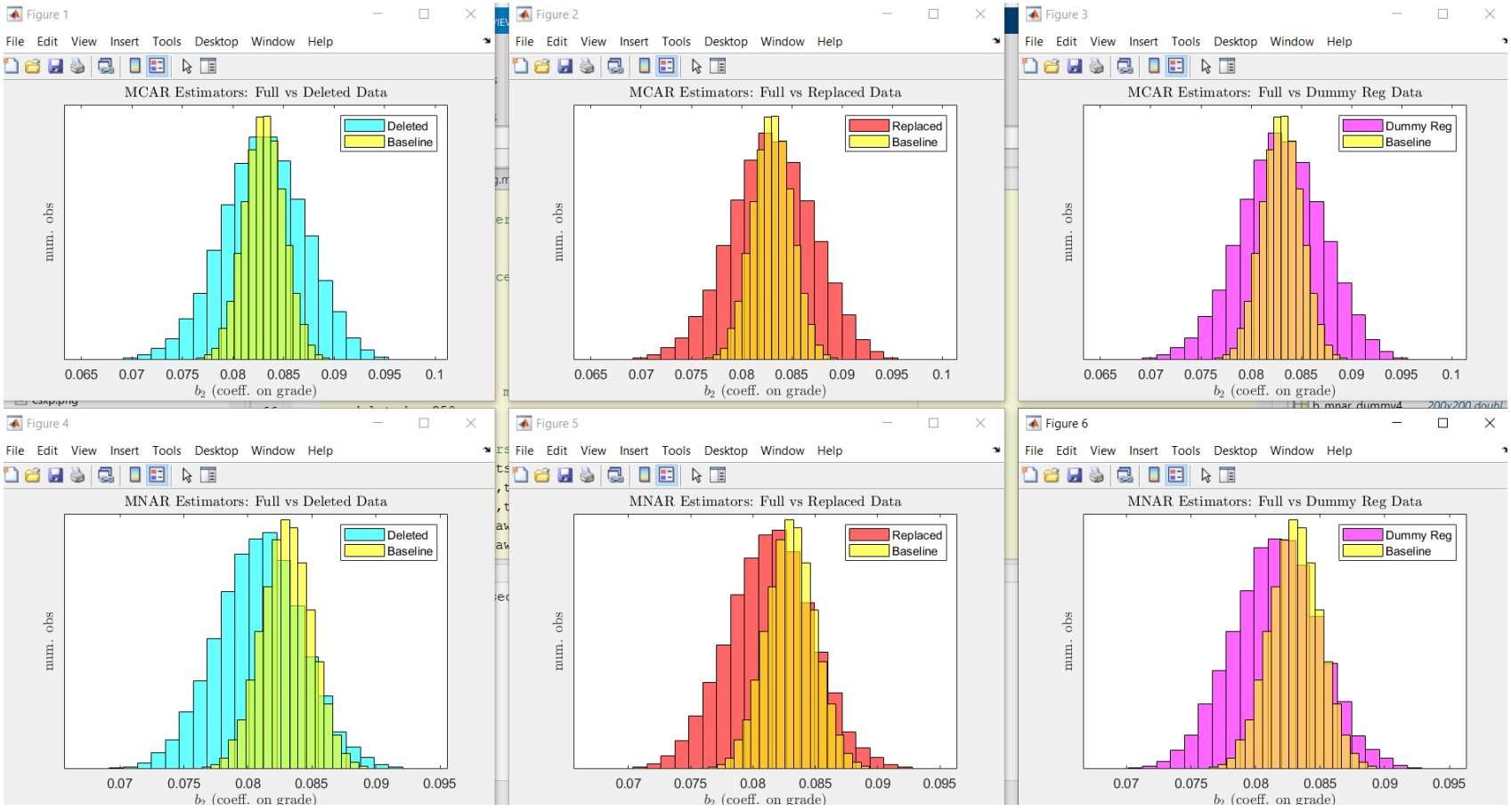






Matlab Graphs

Missing Observations = 250



Matlab Graphs

Missing Observations = 500