Comparison of Imputation Techniques for Missing Data for Potentially Informative Missingness: An Application to Childhood Obesity using the Medical Expenditures Panel Survey

Margie Rosenberg & Michael Wurm

University of Wisconsin – Madison

ARC 2014

Work in Progress

Motivation for the Study

- One-third of US children and teens are either overweight or obese

(Ogden, Carroll, Kit, & Flegal, 2014)

Motivation for the Study

- One-third of US children and teens are either overweight or obese
- Percentage of obese US children aged 6 to 11 years increased from 7% in 1980 to nearly 18% in 2012

(Ogden et al., 2014)

Motivation for the Study

- One-third of US children and teens are either overweight or obese
- Percentage of obese US children aged 6 to 11 years increased from 7% in 1980 to nearly 18% in 2012
- Percentage of obese US adolescents aged 12 to 19 years increased from 5% to nearly 21% over the same period

(Ogden et al., 2014)

Motivation for the Study

- One-third of US children and teens are either overweight or obese
- Percentage of obese US children aged 6 to 11 years increased from 7% in 1980 to nearly 18% in 2012
- Percentage of obese US adolescents aged 12 to 19 years increased from 5% to nearly 21% over the same period
- Long-term population health impact of childhood obesity are greater prevalence of diseases (type 2 diabetes, heart disease), as well as psychological disorders (depression and low self-esteem)

(Ogden et al., 2014)

Definition of Overweight/Obesity in Children

- Body Mass Index (BMI) metric of obesity:

$$\begin{aligned} BMI &= \frac{Kilograms}{Meters^2} \\ &= 703 \cdot \frac{Pounds}{Inches^2} \end{aligned}$$

Definition of Overweight/Obesity in Children

- Body Mass Index (BMI) metric of obesity:

$$\begin{aligned} BMI &= \frac{Kilograms}{Meters^2} \\ &= 703 \cdot \frac{Pounds}{Inches^2} \end{aligned}$$

- *Overweight* = BMI at or above the 85th percentile and lower than the 95th percentile for children of same age and sex

Definition of Overweight/Obesity in Children

- Body Mass Index (BMI) metric of obesity:

$$\begin{aligned} BMI &= \frac{Kilograms}{Meters^2} \\ &= 703 \cdot \frac{Pounds}{Inches^2} \end{aligned}$$

- *Overweight* = BMI at or above the 85th percentile and lower than the 95th percentile for children of same age and sex
- *Obesity*= BMI at or above 95th percentile for children of the same age and sex

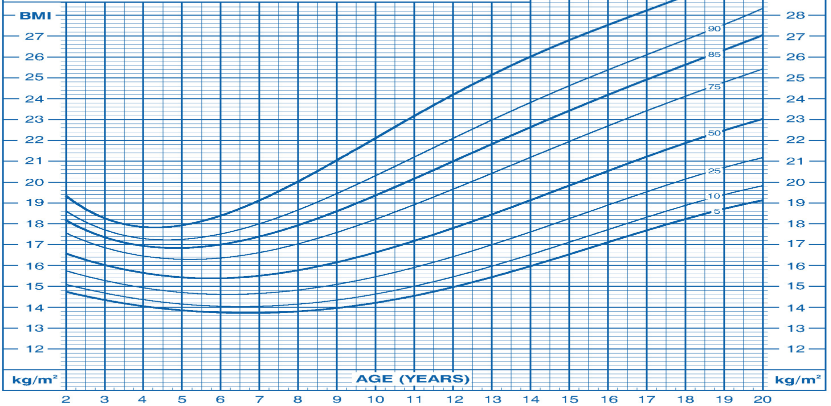
2 to 20 years: Boys
Body mass index-for-age percentiles

NAME ______________________

RECORD # ______________________

Date	Age	Weight	Stature	BMI*	Comments

***To Calculate BMI:** Weight (kg) ÷ Stature (cm) ÷ Stature (cm) x 10,000
or Weight (lb) ÷ Stature (in) ÷ Stature (in) x 703*



Published May 30, 2000 (modified 10/16/00).
SOURCE: Developed by the National Center for Health Statistics in collaboration with
the National Center for Chronic Disease Prevention and Health Promotion (2000).
http://www.cdc.gov/growthcharts

SAFER · HEALTHIER · PEOPLE™

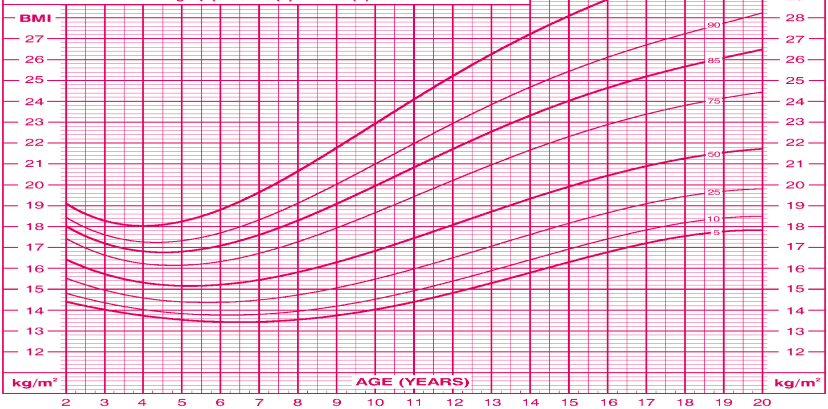
2 to 20 years: Girls
Body mass index-for-age percentiles

NAME ______________________

RECORD # ______________________

Date	Age	Weight	Stature	BMI*	Comments

***To Calculate BMI:** Weight (kg) ÷ Stature (cm) ÷ Stature (cm) x 10,000
or Weight (lb) ÷ Stature (in) ÷ Stature (in) x 703*



Published May 30, 2000 (modified 10/16/00).
SOURCE: Developed by the National Center for Health Statistics in collaboration with
the National Center for Chronic Disease Prevention and Health Promotion (2000).
http://www.cdc.gov/growthcharts

SAFER · HEALTHIER · PEOPLE™

Study Question

To investigate impact of obesity with use of nationally representative, publicly available data, such as the Medical Expenditures Panel Survey (MEPS)

Study Question

To investigate impact of obesity with use of nationally representative, publicly available data, such as the Medical Expenditures Panel Survey (MEPS)

How best to take advantage of MEPS data when interested in BMI as either a dependent or independent variable?

Study Question

To investigate impact of obesity with use of nationally representative, publicly available data, such as the Medical Expenditures Panel Survey (MEPS)

How best to take advantage of MEPS data when interested in BMI as either a dependent or independent variable?

We wish to test the missing data mechanism to best impute information.

Hypothesis

Potentially, data missing with a non-ignorable (or non-informative) mechanism,

Hypothesis

Potentially, data missing with a non-ignorable (or non-informative) mechanism,

Or, data missing at random due to Δ in study questionairre design,

Hypothesis

Potentially, data missing with a non-ignorable (or non-informative) mechanism,

Or, data missing at random due to Δ in study questionairre design,

Or, data missing at random due to child visit pattern to physician,

Hypothesis

Potentially, data missing with a non-ignorable (or non-informative) mechanism,

Or, data missing at random due to Δ in study questionairre design,

Or, data missing at random due to child visit pattern to physician,

Or, data missing at random due to lack of knowledge of parents.

MEPS

- Overall goals of MEPS: To provide unbiased estimates of national and regional (four Census regions) expenditures with a targeted precision, and to provide unbiased estimates for targeted sub-groups, such as race or low income

(McGuire, Glazer, et al., 2013; McGuire, Newhouse, Normand, Shi, & Zuvekas, 2013; S. B. Cohen & Ezzati-Rice, 2006; Ezzati-Rice, Rohde, & Greenblatt, 2009; S. Cohen, 1996; S. B. Cohen, 2000; S. B. Cohen, Ezzati-Rice, Zodet, Machlin, & Yu, 2011; Zuvekas & Olin, 2009)

MEPS

- Overall goals of MEPS: To provide unbiased estimates of national and regional (four Census regions) expenditures with a targeted precision, and to provide unbiased estimates for targeted sub-groups, such as race or low income
- Annual survey of the US civilian, non-institutionalized population (excludes those living in penal, mental, homes for the aged, or members of the armed forces)

(McGuire, Glazer, et al., 2013; McGuire, Newhouse, et al., 2013; S. B. Cohen & Ezzati-Rice, 2006; Ezzati-Rice et al., 2009; S. Cohen, 1996; S. B. Cohen, 2000; S. B. Cohen et al., 2011; Zuvekas & Olin, 2009)

MEPS

- Overall goals of MEPS: To provide unbiased estimates of national and regional (four Census regions) expenditures with a targeted precision, and to provide unbiased estimates for targeted sub-groups, such as race or low income
- Annual survey of the US civilian, non-institutionalized population (excludes those living in penal, mental, homes for the aged, or members of the armed forces)
- Intended to be representative of the US healthcare utilization, and expenditures, insurance

(McGuire, Glazer, et al., 2013; McGuire, Newhouse, et al., 2013; S. B. Cohen & Ezzati-Rice, 2006; Ezzati-Rice et al., 2009; S. Cohen, 1996; S. B. Cohen, 2000; S. B. Cohen et al., 2011; Zuvekas & Olin, 2009)

MEPS

- Overall goals of MEPS: To provide unbiased estimates of national and regional (four Census regions) expenditures with a targeted precision, and to provide unbiased estimates for targeted sub-groups, such as race or low income
- Annual survey of the US civilian, non-institutionalized population (excludes those living in penal, mental, homes for the aged, or members of the armed forces)
- Intended to be representative of the US healthcare utilization, and expenditures, insurance
- Started in 1996 replacing the decennial National Medical Care Expenditure Survey, completed in 1977 and 1987, as a way to provide more timely data for health care expenditures to researchers and policymakers

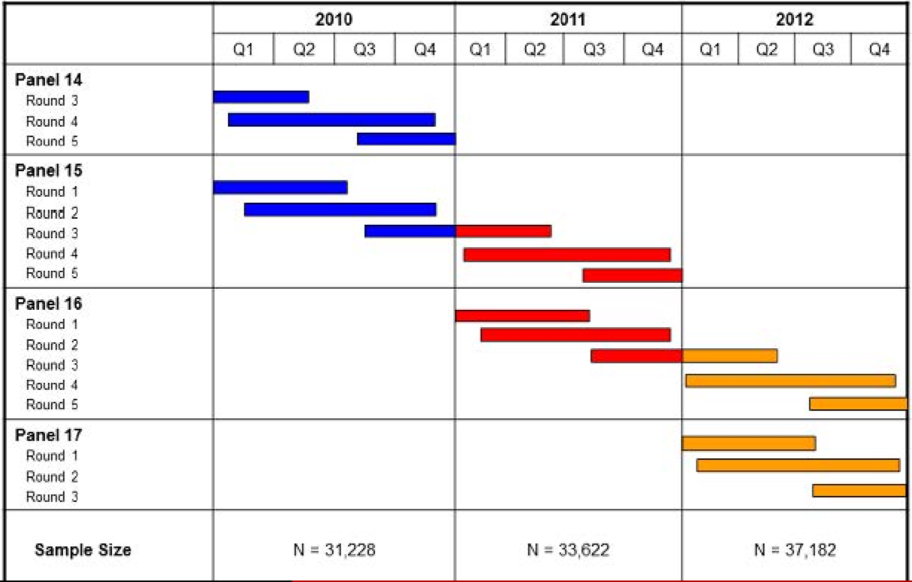
(McGuire, Glazer, et al., 2013; McGuire, Newhouse, et al., 2013; S. B. Cohen & Ezzati-Rice, 2006; Ezzati-Rice et al., 2009; S. Cohen, 1996; S. B. Cohen, 2000; S. B. Cohen et al., 2011; Zuvekas & Olin, 2009)

MEPS

- Overall goals of MEPS: To provide unbiased estimates of national and regional (four Census regions) expenditures with a targeted precision, and to provide unbiased estimates for targeted sub-groups, such as race or low income
- Annual survey of the US civilian, non-institutionalized population (excludes those living in penal, mental, homes for the aged, or members of the armed forces)
- Intended to be representative of the US healthcare utilization, and expenditures, insurance
- Started in 1996 replacing the decennial National Medical Care Expenditure Survey, completed in 1977 and 1987, as a way to provide more timely data for health care expenditures to researchers and policymakers
- MEPS data used in evaluating expenditures for health reform policies and assessing the cost of drugs for Medicare recipients that resulted in adoption of Medicare Part D

(McGuire, Glazer, et al., 2013; McGuire, Newhouse, et al., 2013; S. B. Cohen & Ezzati-Rice, 2006; Ezzati-Rice et al., 2009; S. Cohen, 1996; S. B. Cohen, 2000; S. B. Cohen et al., 2011; Zuvekas & Olin, 2009)

MEPS Longitudinal Design

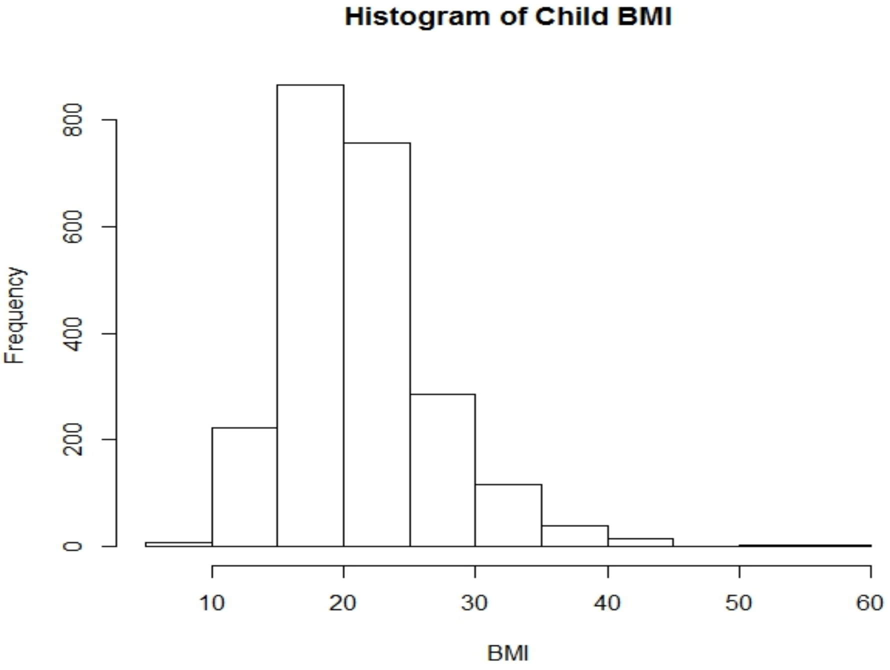


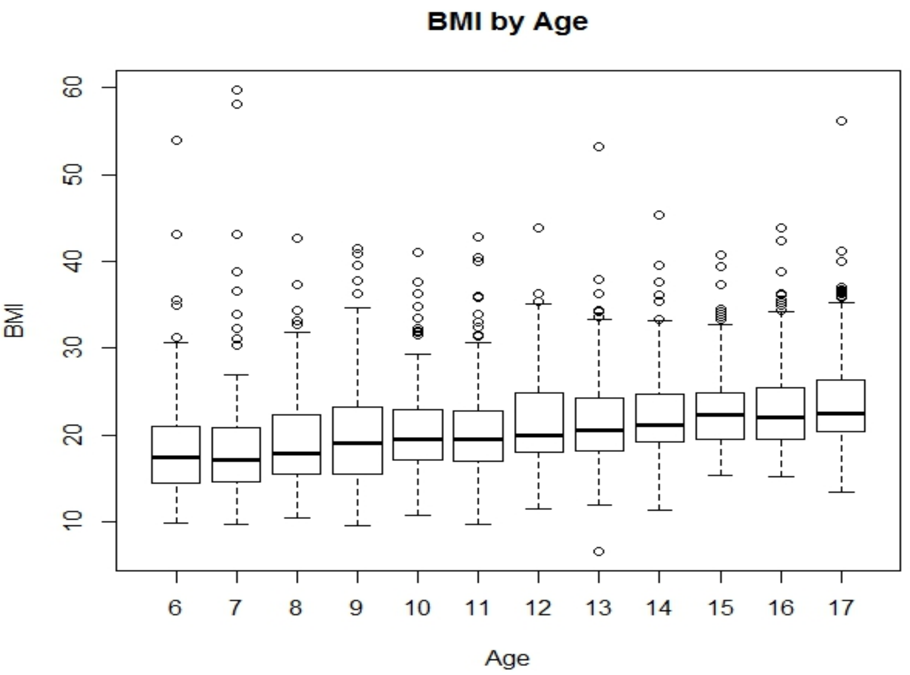
MEPS Data for This Study

- Examine Panel 15 of MEPS in 2011
- Subset children ages 6 to 17 inclusive
- Sample:
  - Total # Observations: 2,855 with almost 20% missing BMI
  - # *Missing* BMI: 539 (537 *true* missing + recoded 2 observations as missing with BMI of 103.3 and 106.2)

Child BMI Descriptives

n	mean	sd	median	min	max	skew	kurtosis	se
2316	21.32	5.77	20.40	6.60	59.80	1.27	3.61	0.12


Histogram of Child BMI
Frequency
0
200
400
600
800
10
20
30
40
50
60
BMI


BMI by Age
BMI
60
50
40
30
20
10
6
7
8
9
10
11
12
13
14
15
16
17
Age

Example: Child Age Descriptives

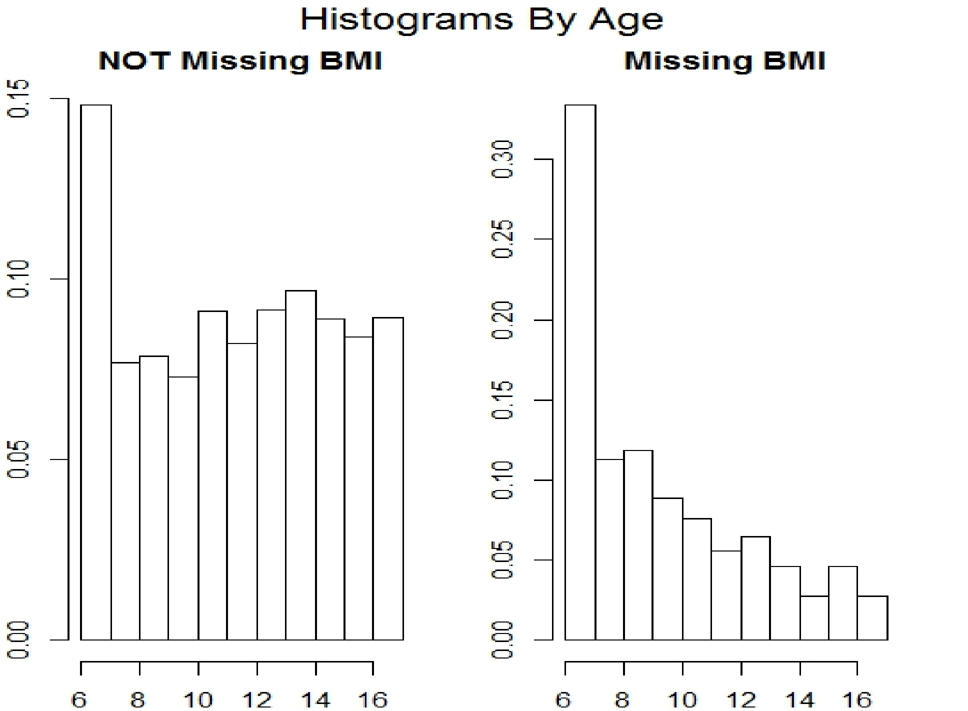
Table: Age Descriptives for Observations with BMI Missing

n	mean	sd	median	min	max	skew	kurtosis	se
539	9.68	3.15	9.00	6.00	17.00	0.71	-0.55	0.14

Table: Age Descriptives for Observations with BMI Not Missing

n	mean	sd	median	min	max	skew	kurtosis	se
2316	11.74	3.41	12.00	6.00	17.00	-0.10	-1.17	0.07

Two-sample t-test of difference between 2 means: p-value < 2.2e-16


Histograms By Age
NOT Missing BMI
Missing BMI
0.00
0.05
0.10
0.15
0.20
0.25
0.30
6
8
10
12
14
16

Missing Data Basics

- Framework of missing data approaches stem to Rubin (1976)

Missing Data Basics

- Framework of missing data approaches stem to Rubin (1976)
- Y_{com} = Complete data response (i.e. hypothetically no missing data)

Missing Data Basics

- Framework of missing data approaches stem to Rubin (1976)
- Y_{com} = Complete data response (i.e. hypothetically no missing data)
- $Y_{com} = (Y_{obs}, Y_{mis})$

Missing Data Basics

- Framework of missing data approaches stem to Rubin (1976)
- Y_{com} = Complete data response (i.e. hypothetically no missing data)
- $Y_{com} = (Y_{obs}, Y_{mis})$
- R = Vector (or matrix) of missingness indicators, with 1 = observed, 0 = missing

Missing Data Basics

- Framework of missing data approaches stem to Rubin (1976)
- Y_{com} = Complete data response (i.e. hypothetically no missing data)
- $Y_{com} = (Y_{obs}, Y_{mis})$
- R = Vector (or matrix) of missingness indicators, with 1 = observed, 0 = missing
- X = Vector/Matrix of auxiliary variables

Missing Data Basics

- Framework of missing data approaches stem to Rubin (1976)
- Y_{com} = Complete data response (i.e. hypothetically no missing data)
- $Y_{com} = (Y_{obs}, Y_{mis})$
- R = Vector (or matrix) of missingness indicators, with 1 = observed, 0 = missing
- X = Vector/Matrix of auxiliary variables
- θ = Vector of parameters

Missing Data Basics

- Framework of missing data approaches stem to Rubin (1976)
- Y_{com} = Complete data response (i.e. hypothetically no missing data)
- $Y_{com} = (Y_{obs}, Y_{mis})$
- R = Vector (or matrix) of missingness indicators, with 1 = observed, 0 = missing
- X = Vector/Matrix of auxiliary variables
- θ = Vector of parameters

Missing Data Basics

- Framework of missing data approaches stem to Rubin (1976)
- Y_{com} = Complete data response (i.e. hypothetically no missing data)
- $Y_{com} = (Y_{obs}, Y_{mis})$
- R = Vector (or matrix) of missingness indicators, with 1 = observed, 0 = missing
- X = Vector/Matrix of auxiliary variables
- θ = Vector of parameters

From:

$$[Y_{com}, R \,|\, X, \theta] \quad \text{or} \quad [Y_{com}, R \,|\, \theta]$$

Missing Data Basics

- Framework of missing data approaches stem to Rubin (1976)
- Y_{com} = Complete data response (i.e. hypothetically no missing data)
- $Y_{com} = (Y_{obs}, Y_{mis})$
- R = Vector (or matrix) of missingness indicators, with 1 = observed, 0 = missing
- X = Vector/Matrix of auxiliary variables
- θ = Vector of parameters

From:

$$[Y_{com}, R \,|\, X, \theta] \quad \text{or} \quad [Y_{com}, R \,|\, \theta]$$

We want to estimate:

$$[Y_{com} \,|\, X, \theta] \quad \text{or} \quad [Y_{com} \,|\, \theta]$$

(Rubin, 1976; Enders, 2010)

Missing Data Mechanism

MCAR: $[R \mid Y_{com}, X, \theta] = [R \mid \theta]$, or observed data points are random sample from complete data

Missing Data Mechanism

MCAR: $[R \mid Y_{com}, X, \theta] = [R \mid \theta]$, or observed data points are random sample from complete data

Missing Data Mechanism

MCAR: $[R \mid Y_{com}, X, \theta] = [R \mid \theta]$, or observed data points are random sample from complete data

MAR: $[R \mid Y_{com}, X, \theta] = [R \mid Y_{obs}, X, \theta]$

Missing Data Mechanism

MCAR: $[R \,|\, Y_{com}, X, \theta] = [R \,|\, \theta]$, or observed data points are random sample from complete data

MAR: $[R \,|\, Y_{com}, X, \theta] = [R \,|\, Y_{obs}, X, \theta]$

Missing Data Mechanism

MCAR: $[R \,|\, Y_{com}, X, \theta] = [R \,|\, \theta]$, or observed data points are random sample from complete data

MAR: $[R \,|\, Y_{com}, X, \theta] = [R \,|\, Y_{obs}, X, \theta]$

MNAR: $[R \,|\, Y_{com}, X, \theta] = [R \,|\, Y_{obs}, Y_{mis}, X, \theta]$

One Approach

- Log(BMI) is approximately normal
- Could just include observed data and carry on
- But we are throwing away 20% of sample

Sampling Distribution vs Likelihood Methods

- $[Y_{com}, \theta]$ has two interpretations:

Sampling Distribution vs Likelihood Methods

- $[Y_{com}, \theta]$ has two interpretations:
  1. Sampling Distribution: $[Y_{com} \mid \theta]$ (non-parametric and semi-parametric procedures)

Sampling Distribution vs Likelihood Methods

- $[Y_{com}, \theta]$ has two interpretations:
  1. Sampling Distribution: $[Y_{com} \mid \theta]$ (non-parametric and semi-parametric procedures)
  2. Likelihood: $[\theta \mid Y_{com}]$ (Maximum likelihood and multiple imputation)

Sampling Distribution vs Likelihood Methods

- $[Y_{com}, \theta]$ has two interpretations:
  1. Sampling Distribution: $[Y_{com} \,|\, \theta]$ (non-parametric and semi-parametric procedures)
  2. Likelihood: $[\theta \,|\, Y_{com}]$ (Maximum likelihood and multiple imputation)
- We could focus only on Observed Data by integrating out missing data:

Sampling Distribution vs Likelihood Methods

- $[Y_{com}, \theta]$ has two interpretations:
  1. Sampling Distribution: $[Y_{com} \,|\, \theta]$ (non-parametric and semi-parametric procedures)
  2. Likelihood: $[\theta \,|\, Y_{com}]$ (Maximum likelihood and multiple imputation)
- We could focus only on Observed Data by integrating out missing data:

Sampling Distribution vs Likelihood Methods

- $[Y_{com}, \theta]$ has two interpretations:
  1. Sampling Distribution: $[Y_{com} \,|\, \theta]$ (non-parametric and semi-parametric procedures)
  2. Likelihood: $[\theta \,|\, Y_{com}]$ (Maximum likelihood and multiple imputation)
- We could focus only on Observed Data by integrating out missing data:

$$[Y_{obs}, \theta] = \int [Y_{com}, \theta] \, dY_{mis} \tag{1}$$

- For (1) to be correct sampling distribution, need missing data to be MCAR, otherwise results are biased

(Rubin, 1976; Schafer & Graham, 2002)

Sampling Distribution vs Likelihood Methods

- $[Y_{com}, \theta]$ has two interpretations:
  1. Sampling Distribution: $[Y_{com} \,|\, \theta]$ (non-parametric and semi-parametric procedures)
  2. Likelihood: $[\theta \,|\, Y_{com}]$ (Maximum likelihood and multiple imputation)
- We could focus only on Observed Data by integrating out missing data:

$$[Y_{obs}, \theta] = \int [Y_{com}, \theta] \, dY_{mis} \tag{1}$$

- For (1) to be correct sampling distribution, need missing data to be MCAR, otherwise results are biased
- For (1) to be correct likelihood, need missing data to be MAR

(Rubin, 1976; Schafer & Graham, 2002)

Our Results So Far

- Compare methods that assume MAR

Our Results So Far

- Compare methods that assume MAR
  1. Stochastic Regression: Impute $\widehat{log(BMI)} + \varepsilon$, where $\varepsilon \sim Normal(0, \widehat{\sigma}^2)$

Our Results So Far

- Compare methods that assume MAR
  1. Stochastic Regression: Impute $\widehat{log(BMI)} + \varepsilon$, where $\varepsilon \sim Normal(0, \widehat{\sigma}^2)$
  2. Semi-Parametric Stochastic Regression: Impute $\widehat{log(BMI)} + r$, where r drawn from residuals

Our Results So Far

- Compare methods that assume MAR
  1. Stochastic Regression: Impute $\widehat{log(BMI)} + \varepsilon$, where $\varepsilon \sim Normal(0, \widehat{\sigma}^2)$
  2. Semi-Parametric Stochastic Regression: Impute $\widehat{log(BMI)} + r$, where r drawn from residuals
  3. Hot Deck: Randomly select an observed value from the pool of observations that match on selected covariates

Our Results So Far

- Compare methods that assume MAR
  1. Stochastic Regression: Impute $\widehat{log(BMI)} + \varepsilon$, where $\varepsilon \sim Normal(0, \widehat{\sigma}^2)$
  2. Semi-Parametric Stochastic Regression: Impute $\widehat{log(BMI)} + r$, where r drawn from residuals
  3. Hot Deck: Randomly select an observed value from the pool of observations that match on selected covariates
- Approach:

Our Results So Far

- Compare methods that assume MAR
  1. Stochastic Regression: Impute $\widehat{log(BMI)} + \varepsilon$, where $\varepsilon \sim Normal(0, \widehat{\sigma}^2)$
  2. Semi-Parametric Stochastic Regression: Impute $\widehat{log(BMI)} + r$, where r drawn from residuals
  3. Hot Deck: Randomly select an observed value from the pool of observations that match on selected covariates
- Approach:
  1. Consider complete BMI cases as "truth"

Our Results So Far

- Compare methods that assume MAR
  1. Stochastic Regression: Impute $\widehat{log(BMI)} + \varepsilon$, where $\varepsilon \sim Normal(0, \widehat{\sigma}^2)$
  2. Semi-Parametric Stochastic Regression: Impute $\widehat{log(BMI)} + r$, where r drawn from residuals
  3. Hot Deck: Randomly select an observed value from the pool of observations that match on selected covariates
- Approach:
  1. Consider complete BMI cases as "truth"
  2. Randomly select 20% as missing

Our Results So Far

- Compare methods that assume MAR
  1. Stochastic Regression: Impute $\widehat{log(BMI)} + \varepsilon$, where $\varepsilon \sim Normal(0, \widehat{\sigma}^2)$
  2. Semi-Parametric Stochastic Regression: Impute $\widehat{log(BMI)} + r$, where r drawn from residuals
  3. Hot Deck: Randomly select an observed value from the pool of observations that match on selected covariates
- Approach:
  1. Consider complete BMI cases as "truth"
  2. Randomly select 20% as missing
  3. For each method, generate 1,000 sets of imputed data

Our Results So Far

- Compare methods that assume MAR
  1. Stochastic Regression: Impute $\widehat{log(BMI)} + \varepsilon$, where $\varepsilon \sim Normal(0, \widehat{\sigma}^2)$
  2. Semi-Parametric Stochastic Regression: Impute $\widehat{log(BMI)} + r$, where r drawn from residuals
  3. Hot Deck: Randomly select an observed value from the pool of observations that match on selected covariates
- Approach:
  1. Consider complete BMI cases as "truth"
  2. Randomly select 20% as missing
  3. For each method, generate 1,000 sets of imputed data
  4. Compare imputed distribution of missing value to true distribution

Results: Using Log(BMI)

Approach	Mean	Std. Dev.	MSPE
Stochastic Regression	3.021	0.258	0.0668
Semi-Parametric Regression	3.022	0.259	0.0672
Hot Deck	3.018	0.260	0.0687
Truth	3.051	0.273	

Future Work

- Explore MNAR approaches
- Choose informative missing pattern for data from MEPS and test models

Bibliography I

Cohen, S. (1996). The redesign of the medical expenditure panel survey: A component of the dhhs survey integration plan. In *Proceedings of the copafs seminar on statistical methodology in the public service.*

Cohen, S. B. (2000). *Sample design of the 1997 medical expenditure panel survey, household component* (No. 1). US Department of Health and Human Services, Public Health Service, Agency for Healthcare Research and Quality.

Cohen, S. B., & Ezzati-Rice, T. M. (2006). Designing national health care surveys to inform health policy. In (p. 89-101). Thomson, Brooks/Cole.

Cohen, S. B., Ezzati-Rice, T. M., Zodet, M., Machlin, S., & Yu, W. (2011). An assessment of the impact of two distinct survey design modifications on health care utilization estimates in the medical expenditure panel survey. *Journal of Economic and Social Measurement*, *36*(1), 33–69.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Ezzati-Rice, T. M., Rohde, F., & Greenblatt, J. (2009). *Methodology report no. 22: sample design of the medical expenditure panel survey household component, 1998-2007.*

Bibliography II

McGuire, T. G., Glazer, J., Newhouse, J. P., Normand, S.-L., Shi, J., Sinaiko, A. D., & Zuvekas, S. H. (2013). Integrating risk adjustment and enrollee premiums in health plan payment. *Journal of health economics*, *32*(6), 1263–1277.

McGuire, T. G., Newhouse, J. P., Normand, S.-L., Shi, J., & Zuvekas, S. (2013). Assessing incentives for service-level selection in private health insurance exchanges. *Journal of Health Economics (Under Revision)*.

Ogden, C. L., Carroll, M. D., Kit, B. K., & Flegal, K. M. (2014). Prevalence of childhood and adult obesity in the united states, 2011-2012. *JAMA*, *311*(8), 806–814.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, *7*(2), 147.

Zuvekas, S. H., & Olin, G. L. (2009). Validating household reports of health care use in the medical expenditure panel survey. *Health services research*, *44*(5p1), 1679–1700.