Approaches to Missing Data

A Presentation by
Russell Barbour, Ph.D.
Center for Interdisciplinary Research on AIDS (CIRA)
and
Eugenia Buta, Ph.D.
CIRA and The Yale Center of Analytical Studies (YCAS)
April 15, 2013

Purpose

- A brief description of modeling missing data using a Bayesian approach via ***WinBUGS*** software

- Focus on multiple imputation methods available in ***R***, in particular the ***Amelia*** package which incorporates some aspects of Bayesian modeling

Key Elements of Missingness

- The number of cases missing per variable
- The number of variables missing per case
- The pattern of correlations among variables

Types of Missing Data
(from McKnight et al 2007)

- **The sampling process involves both the selection of the units, and the process by which observations become missing - the *missingness mechanism***

- **For valid inference, we need to take account of the missingness mechanism**

Categorizing Missingness

- Missing completely at Random *MCAR*

  No difference between the characteristics of those observed and those missing---***tested with Little's MCAR test***

  Missingness not related to the data values (e.g. lab or data entry mistakes, subject drops out of the study for reason not related to intervention)

- Missing at Random *MAR*

  There is a difference, but it can be explained by other variables or data

  Missingness depends only on observed data "e.g. subjects with missing income younger, on average, than those reporting income"

- Missing Not at Random *MNAR*

  The difference cannot be explained; missingness is directly related to the unobserved data (e.g. probability of reporting income depends on level of income; in a clinical trial, subjects drop out because too sick)---"a worse case scenario"

Missing Data Consequences

Bias

- Estimate systematically deviates from the quantity of interest
- No bias if data is MCAR, but bias can occur with not MCAR

Variance

- Missing data can sometimes lead to wrong standard errors
- Wrong study conclusions about relationship of variables to outcomes

Impact of missing data depends on amount of missing data and missing mechanism

MNAR Consequences

- MNAR means that you cannot simply remove the missingness and do the analysis, even if you seem to have a large N number. (true of MAR too but as we shall see this situation can be mitigated)

- This is because the results of such an analysis would be biased

- For example in a study of early or late treatment of HIV+ children we had to ascertain if the missing children were still alive

- Difficult to detect at times since you may not have the "known unknowns" to make the judgment

- Distinguishing MAR from MNAR only from observed data is difficult: we cannot verify the MNAR without knowing the values of the missing data

Little's MCAR test

- Outlined in :

  Little RJA (1988) A Test of Missing Completely at Random for Multivariate Data with Missing Values. Journal of the American Statistical Association 83: 1198-1202.

- *The test is based on mean differences across groups of subjects with the same missing data pattern "*

- *Since the power [of the test] may be low, it is prudent to keep in mind that accepting the null hypothesis of MCAR does not imply its correctness" (Little , 1988)*

- Implemented in SAS

- Also available in SPSS but as part of a costly missing data add–on

- IRM core at CIRA will do Little's test for affiliated research projects. Both Eugenia and I have run SAS code successfully

An Example of Little's MCAR

- The SAS System 16:09 Friday, July 27, 2012 1
- Number of Observed Variables = 13
- Number of Missing Data Patterns = 8
- Summary of Missing Data Patterns (0 = Missing, 1 = Observed)
- Frequency | Pattern | d2j
- 1 | 1 1 1 1 0 1 1 1 1 1 1 1 1 | 13.75164
- 4 | 1 1 1 1 1 1 0 0 0 0 0 0 0 | 14.02515
- 42 | 1 1 1 1 1 1 0 0 0 0 0 0 1 | 7.603233
- 1 | 1 1 1 1 1 1 0 1 1 1 1 1 0 | 7.651030
- 2 | 1 1 1 1 1 1 1 0 1 1 1 1 0 | 8.979970
- 2 | 1 1 1 1 1 1 1 0 1 1 1 1 1 | 12.89867
- 23 | 1 1 1 1 1 1 1 1 1 1 1 1 0 | 9.255660
- 73 | 1 1 1 1 1 1 1 1 1 1 1 1 1 | 7.840028
- Sum of the Number of Observed Variables Across Patterns (Sigma psubj) = 84
- Little's (1988) Chi-Square Test of MCAR
- Chi-Square (d2) = 82.005
- df (Sigma psubj - p) = 71
- **p-value = 0.175**

Explanation of Little's test in Shenoi et al 2012

• Conducted the Little's test on individual four factors used to create the LCSI using open code in ***SAS.***

• Results of these four tests determined that missingness for hemoglobin, albumin, and alkaline phosphatase were missing completely at random relative to survival.

•Although creatinine was found to be significantly related to survival (p = .000), when Little's test was applied to a combined analysis for all four components of the LCSI in relationship the congregate analysis demonstrated that taken together the data is missing completely at random.

•As noted by McKnight et al 2007, Little's test done with a combination of all variables is the more robust test when compared with individual tests where alpha inflation distorts results.

Little’s MCAR Test Values in relationship to Survival (alive after 180 days)

Little's Test	*P*
AlkPhos	ns
CREA	0.000
Hemo	ns
Album	ns
All4	ns

Methods That Risk Bias

- **Complete case analysis**: analyze only subjects with complete data—may waste data and lose power
- **Single imputation**: fill-in missing values by the mean of the observed values and then treat the data as complete, or fill-in by predicted value obtained from fitting a regression model to the observed values---**variability typically underestimated**: we treat filled-in values as real (observed) data and do not take into consideration uncertainty surrounding them
- **Missing data indicator:** creating separate "missing" category for categorical variables—can lead to bias
- **Last Observation Carried Forward:** distorts means/variances

Last Observation Carried Forward: Example

	BMI time1	BMI time2	BMI time3	BMI time4
Subject 1	25.5	26.1	27	? 27
Subject 2	24	? 24	? 24	? 24

Missing Data: Imputation versus Modeling

(from Little and Rubin 2002)

- Bayesian Algorithms for Missing data are *Model Based* methods, in contrast to older *Imputations Based Methods*

- For many years the computationally simpler imputation based methods were used, but such methods are highly biased, mean imputation, regression, " hot deck " etc

Missing Data: Model Based Approach

- A model for the data is defined and which results in the likelihood or posterior distribution of the model
- Estimates in general including those of the variance can be formed to account for the missing data
- This is the basis of modeling missing data in *WinBUGS*

What is WinBUGS?

- Joint endeavor between the MRC Biostatistics Unit in Cambridge and the Department of Epidemiology and Public Health of Imperial College, at St Mary's Hospital London

- BUGS: **B**ayesian Inference **U**sing **G**ibbs **S**ampling

- WinBUGS: DoodleBUGS (graphical representation of model) and window interface for controlling the analysis

What is WinBUGS?

- WinBUGS can handle analysis for complex statistical models
  - Missing data
  - Measurement error
  - No closed form for posterior distribution
- Bayesian posterior inference achieved via Markov chain Monte Carlo (MCMC) Integration
  - Useful when no closed form exists.

WinBUGS is Not for the Faint of Heart!!

In our example I applied non-informative priors "Half-Normal"

Simulations set at 50,000 iterations

" Burn in" = 4000

Sample Code

```
# MODEL – start
{
#############
### PRIOR ###
#############
Beta0 ~ dnorm(0, 1.00E-4)

### X Effects: independent ("Fixed"), corner constraints via tiny priorSD.X[] ###
### priorSD.X[] read from 'BGX DataList.txt'
for (i in 1: N.X){
    for (j in 1: levs.X[i]){

        priorInvV.X[i, j] <- pow(priorSD.X[j, i], -2)
        X.Eff[i, j] ~ dnorm(0, priorInvV.X[i, j])
    }
}

### Z Effects: assumed exchangeable ('Random') ###
### Half-N prior: hyper.Z is SD of prior for Z.Eff[] ###
for (j in 1: 6){
    Z.Eff[1, j] ~ dnorm(0, tau.Z[1])
}
tau.Z[1] <- pow(sigma.Z[1], -2)
sigma.Z[1] <- hyper.Z[1] + 1.E-20
hyper.Z[1] ~ dnorm(0, 0.04)I(0,)
### Half-N prior: hyper.Z is SD of prior for Z.Eff[] ###
for (j in 1: 3){
    Z.Eff[2, j] ~ dnorm(0, tau.Z[2])
}
tau.Z[2] <- pow(sigma.Z[2], -2)
sigma.Z[2] <- hyper.Z[2] + 1.E-20
hyper.Z[2] ~ dnorm(0, 0.04)I(0,)
### Half-N prior: hyper.Z is SD of prior for Z.Eff[] ###
for (j in 1: 3){
    Z.Eff[3, j] ~ dnorm(0, tau.Z[3])
}
tau.Z[3] <- pow(sigma.Z[3], -2)
sigma.Z[3] <- hyper.Z[3] + 1.E-20
hyper.Z[3] ~ dnorm(0, 0.04)I(0,)
```

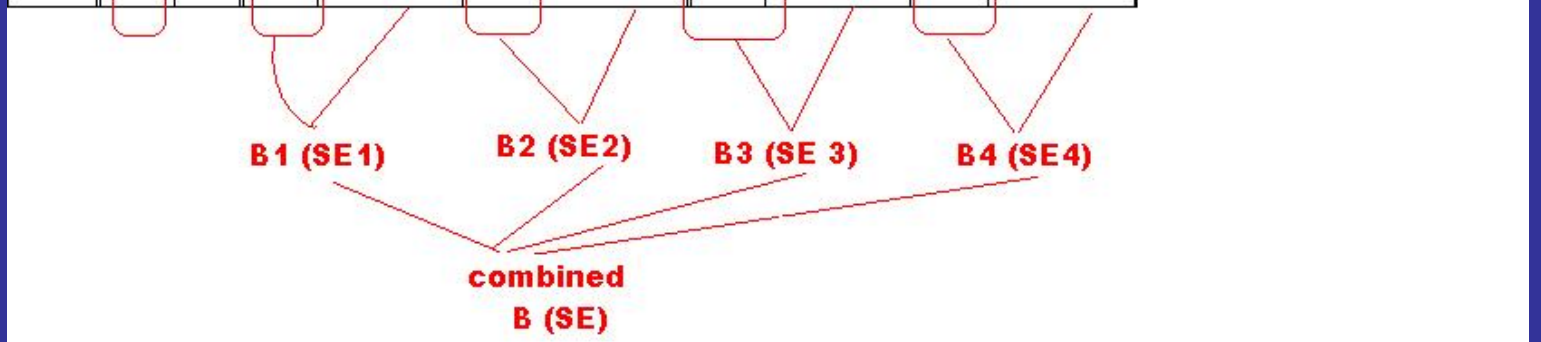
```
        ###################
        ### LIKELIHOOD ###
        ###################
        for (j in 1: N.obs){
            for (i in 1: N.X){
                X.row[i, j] <- X.Eff[i, X[j, i]]
            }
            for (i in 1: N.Z){
                Z.row[i, j] <- Z.Eff[i, Z[j, i]]
            }

            logit(mu[j]) <- Beta0 + sum(X.row[, j]) + sum(Z.row[, j])
            Y[j] ~ dbin(mu[j], N[j])
        }
        ################################
        ### PREDICTIONS & CONTRASTS ###
        ################################
        Pred.Xrow[1, 1] <- X.Eff[1, 1]
        Pred.Xrow[1, 2] <- X.Eff[1, 2]
        Pred.Xrow[1, 3] <- X.Eff[1, 3]
        Pred.Xrow[1, 4] <- X.Eff[1, 4]
        Pred.Xrow[1, 5] <- X.Eff[1, 5]
        Pred.Zrow[2, 2] <- 0
        ……..
        Pred.Zrow[3, 6] <- 0
        Pred.Beta0[1] <- Beta0
        Pred.Beta0[2] <- Beta0
        Pred.Beta0[3] <- Beta0
        Pred.Beta0[4] <- Beta0
        Pred.Beta0[5] <- Beta0
        Pred.Beta0[6] <- Beta0
        for (j in 1: N.pred){
            Pred.Odds[j] <- exp(Pred.Beta0[j] + sum(Pred.Xrow[, j]) + sum(Pred.Zrow[, j]))
            Pred.Ave[j] <- Pred.Odds[j] / (1 + Pred.Odds[j])
        }

    # MODEL – end
        *** KEY ***
        Y[] -> hyper_dx/count
        X.Eff[1,] -> Physical_Demand

        Z.Eff[1,] -> agebin
        Z.Eff[2,] -> BMIbin
        Z.Eff[3,] -> currsmoke
```

Missing Data: Multiple Imputation

- perform imputation of the missing data M times (usually 5 or more)
- the M complete data sets are analyzed separately and then results across the M analyses are combined
- assumes MAR

An Example of Multiple Imputation

	Data		Imputation 1		Imputation 2		Imputation 3		Imputation 4	
Unit	Y_M	Y_O	Y_M	Y_O	Y_M	Y_O	Y_M	Y_O	Y_M	Y_O
1	1.1	3.4	1.1	3.4	1.1	3.4	1.1	3.4	1.1	3.4
2	1.5	3.9	1.5	3.9	1.5	3.9	1.5	3.9	1.5	3.9
3	2.3	2.6	2.3	2.6	2.3	2.6	2.3	2.6	2.3	2.6
4	3.6	1.9	3.6	1.9	3.6	1.9	3.6	1.9	3.6	1.9
5	0.8	2.2	0.8	2.2	0.8	2.2	0.8	2.2	0.8	2.2
6	3.6	3.3	3.6	3.3	3.6	3.3	3.6	3.3	3.6	3.3
7	3.8	1.7	3.8	1.7	3.8	1.7	3.8	1.7	3.8	1.7
8	?	0.8	**0.2**	0.8	**0.8**	0.8	**0.3**	0.8	**2.3**	0.8
9	?	2.0	**1.7**	2.0	**2.4**	2.0	**1.8**	2.0	**3.5**	2.0
10	?	3.2	**2.7**	3.2	**2.5**	3.2	**1.0**	3.2	**1.7**	3.2



Plan B: ***AMELIA*** a package in ***R*** for multiple imputation

Indeed named for Amelia Earhart, a symbol of “missingness”

What is *R* ?

- It is both a computer programming language and a software package.
- Derived from ***S*** programming language created by John Chambers while at Bell Labs
- Current R version was developed in New Zealand by Ross Ihaka and Robert Gentleman
- It is both Freeware and open source…. i.e. no charge and the underlying code is available to anyone

What is *R* ?

- *de facto* standard among statisticians for the development of statistical software
- Along with the standard software, there are almost 3000 specialized task-specific add-on packages for almost any statistical analysis you might want to try

In this presentation we will use the ***Amelia*** add-on package developed by Honaker, King and Blackwell as revised on April 3, 2013

What Does Amelia Do?

- Resamples the original data set using a bootstrap algorithm
- Implements an expectation–maximization (EM) algorithm
- An expectation–maximization (EM) algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models
- Uses all of the data to estimate missing values

Multiple Imputation?

- If you think about it, the sample even with a large amount of " missingness" is but one realization of the data…" a possible reality"
- Amelia creates other realizations of the data using the aforementioned algorithms
- Thus it creates "other possible realities"
- The default in Amelia is set to 5...How many realities do you need?
- All five are combined to apply standard statistical analysis and further reduce bias (using the whimsically named "***Zelig***" algorithms in ***R"***

Capabilities of Amelia II

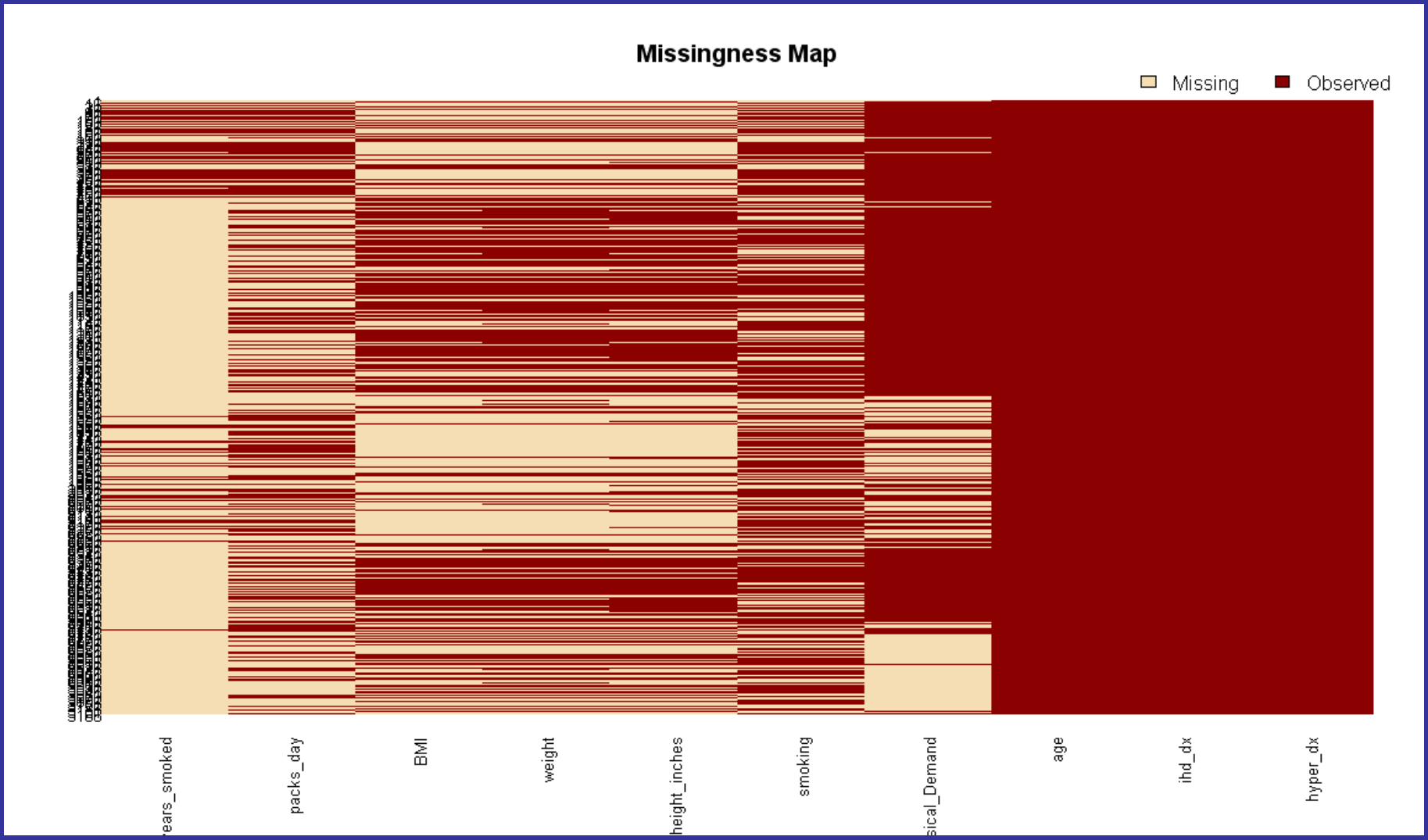
- ‘‘Multiply imputes’’ missing data in a single cross-section (such as a survey), from a time series (like variables collected for each year in a country), or from a time-series-cross-sectional data set (such as collected by years for each of several countries)

- Implements a bootstrapping-based algorithm that gives essentially the same answers as the standard approaches, but faster and can handle many more variables.

- Generalizes existing approaches by allowing for trends in time series across observations within a cross-sectional unit , as well as priors that allow experts to incorporate beliefs they have about the values of missing cells in their data.

- GUI makes detailed knowledge of ***R*** unnecessary

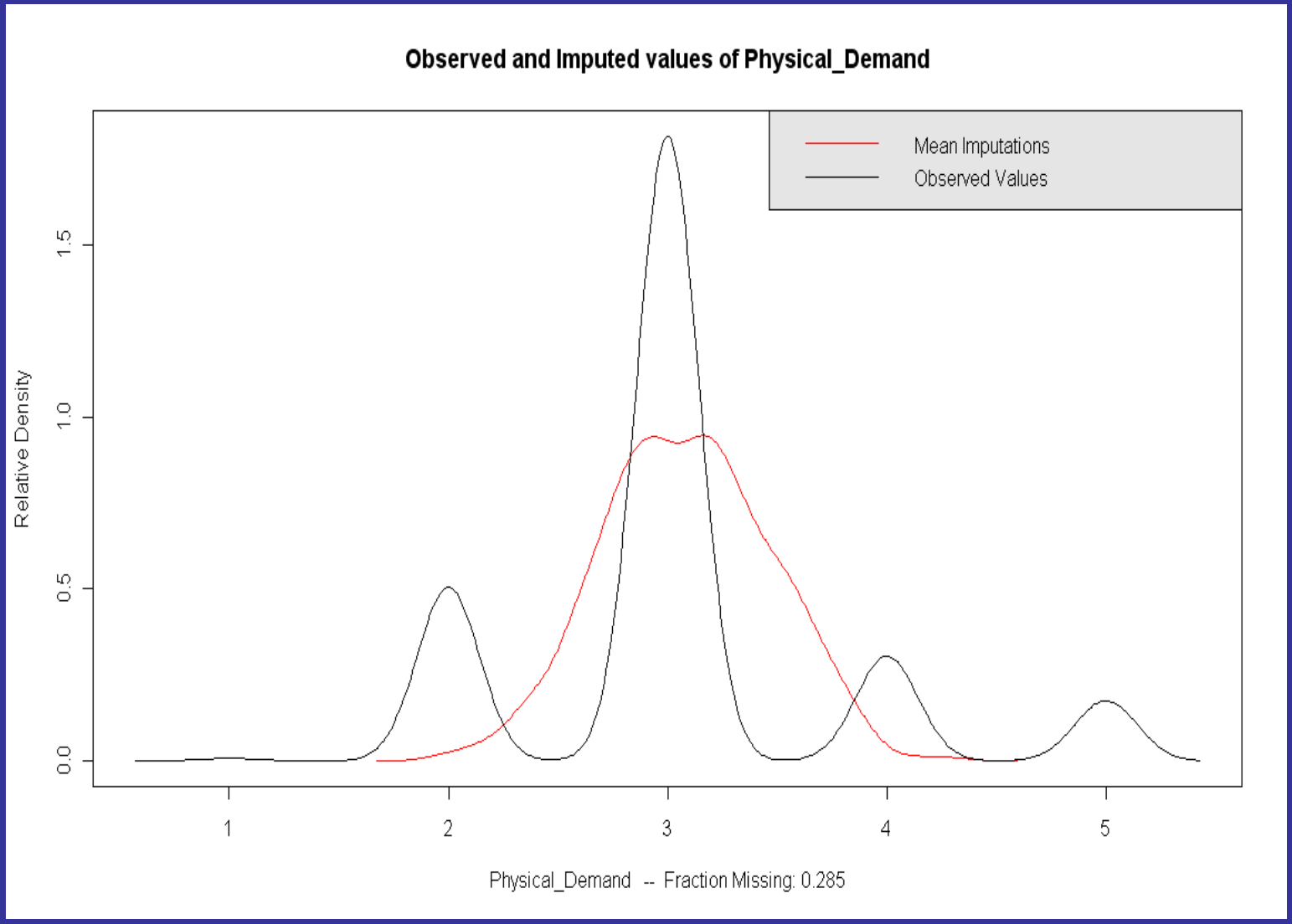
Amelia Will Map the Missingness for you



Elegantly Simple Coding

a.out <-amelia(x = Occup)

The process is not perfect !!



IRM Core Services

- Application of Little's Test
- Bayesian simulations
- Exploration of Missingness (see how the observations with missingness differ from data that is complete)
- Instruction and/or implementation of Multiple Imputation in Amelia II (Barbour) and SAS (Buta)

Thank you for your attention!!

What Amelia Does Exactly

- The model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which computes the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step
- Then computes parameters maximizing the expected log-likelihood found on the *E* step
- These parameter-estimates are then used to determine the distribution of the latent variables in the next E step