W4 Imputing for Missing Items

Reasons for Imputation

The Missingness Problem
• All nonsample cases are completely missing. We impute for those by assigning weights to sample cases
• Missing items in otherwise complete responses
  - special codes used for missing items:
    - NA in R
    - .a. to .z ., and ._ in SAS
    - .a to .z and . in STATA
      Particular surveys may use their own codes for missing: 99, -9
• Ways of handling
  - complete case analysis (casewise deletion)
  - available case analysis – similar to complete case
    - A case is deleted if it is missing on any variables in a particular analysis
  - impute for missing values

Problems with complete case analysis
• If units with missing values differ systematically from completely observed cases, this could bias the complete-case analysis
• If many variables are included in a model, there may be very few complete cases ➔ most of data discarded for sake of a simpler analysis
• Dropped cases not really ignored
  - implied imputation: every missing case is imputed by the average of the complete cases

Missing Data Mechanisms
• MCAR – every unit has same probability of appearing in a sample
• MAR – probability of appearing depends on covariates known for sample and nonsample cases
• NINR – probability of appearing depends on covariates and y's

MAR is usually the best we hope for – if enough covariates are used in imputations, the we avoid NINR

Means and Hotdeck

Methods of imputation
[1] imputation based on logical rules – infer a value based on answers to other questions (these are more like edit checks)
[2] Mean (usually within cells) with or without random error
[3] cold deck (last value carried forward in longitudinal survey)
[4] hot deck – impute value from a similar complete case
[5] regression with or without random error
[6] predictive mean matching – find unit with closest observed value to one predicted by regression

Each method can be done sequentially where item with fewest missing values is imputed first
Then, those imputations + complete values are used to impute item with next-most missing etc

Mean imputation
• If many missing values, a spike is introduced in distribution of a variable
• Random error added to mean reduces distortion. Normal error with mean 0 and variance equal to observed element variance of nonmissing values. Distributions other than normal can be used.
• Cells or subgroups are way of accounting for possibility that value depends on covariates
• Special case of regression imputation

Hot deck imputation
• Put units into groups
  - type of business x size
  - age x gender
• If a unit is missing an item, select a value at random from the cases that reported that item
• Implicit assumption is that all units in a group have a common mean

Regression

Regression imputation with a random error

- Continuous variables
- Regression model fit with complete data

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i$$

- If case k is missing y, the imputation is

$$\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_{1k} + \hat{\beta}_2 x_{2k} + \cdots + \hat{\beta}_p x_{pk} + \epsilon_k^*$$

where $\hat{\beta}_j$'s are estimates based on complete cases

- ϵ_k^* can be a random draw from the set of sample residuals for the complete cases
- Case k must have all of the x's present
- If covariates include main effects and all interactions of a set of categorical variables, this is mean imputation with a random error added

Predictive Mean Matching
• use complete data to regress y on x's
• predict mean for a case with missing y based on the regression (need complete covariates for a missing case)
• Find respondent whose observed value is closest to predicted mean
• Impute that respondent's value to the missing case
➔ ppm is more flexible than hot deck in allowing covariates to be used in the imputation
➔ pmm allows both main effects and interactions of categorical variables to be used in the model

Discrete Variables

- Assign $y_i = 1$ if case i has a characteristic, 0 if not
- Logistic (or other binary) regression model fit with complete data

$$logit(p_k) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$$

- If case k is missing y, the imputed mean on the logit scale is

$$\hat{z}_k = \hat{\beta}_0 + \hat{\beta}_1 x_{1k} + \hat{\beta}_2 x_{2k} + \cdots + \hat{\beta}_p x_{pk}$$

- Back-transform to probability scale with

$$\hat{p}_k = \frac{exp(\hat{z}_k)}{1 + exp(\hat{z}_k)}$$

- Generate uniform random number u in [0,1]
- If $u \leq \hat{p}_k$, $y = 1$; if not $y = 0$

Effect on Variances of imputing for missing values

Imputation variance
• Imputations add variance to estimates that should be accounted for
• Treating imputations as real data will lead to SEs that are too small for most estimates
• Various ways of doing this (and theoretical arguments to justify methods)
 - some require specialized formulas that depend on how imputations were made
 - multiple imputation is more general but requires that there be randomness in how imputations are generated

Multiple Imputation
• Multiple imputation (MI) is one way of reflecting extra variance due to item imputations
 - impute more than one value (m) for each missing item on each case
 - must be some random element in imputation to allow this
 - use special formula to account for imputation variance

var(estimator with imputations) =

(variance treating imputations as real) +

(average variance between estimates using different imputed values)

- Compute estimate Q_t from each of $t = 1, \ldots, m$ completed datasets (complete means real collected values + imputed values for missing variables)
- Estimate for the item is $\bar{Q} = m^{-1} \sum_{t=1}^{m} Q_t$
- Variance estimates U_t treating all data as complete (not imputed)
  - U_t can be any variance estimate appropriate for sample design and estimator
  - If design is $stsrs$ and estimator is the mean, $Q_t = \bar{y}_t$ use

$$U_t \equiv v(\bar{y}_t) = \sum_{h=1}^{H} W_h^2 (1 - f_h) s_{ht}^2 / n_h$$

- Variance of $\bar{Q}$

$$v(\bar{Q}) = \bar{U} + (1 + m^{-1}) B$$

with $\bar{U} = m^{-1} \sum_{t=1}^{m} U_t$ and

$B = (m - 1)^{-1} \sum_{t=1}^{m} (Q_t - \bar{Q})^2$

- Note that MI is a method of variance estimation **not** a method of imputation
- The MI variance formula applies to any method in which the imputed value is random
  - Mean imputation with random error
  - Hot deck with random draws from completes
  - Regression with random error

Pros and Cons of multiple imputation
• Advantages
  - simple variance formula
  - same variance formula applies for many types of estimates (e.g. means, totals, quantiles)
  - point estimates and variance estimates of point estimates are approximately unbiased if imputation model is correct
  - uses all available data – no cases discarded
• Disadvantages
  - MI variance estimator can be positively biased in some cluster samples

Software for Imputation

Software
- mi, mice package in R
- IVEware SAS macro
- proc mi in SAS
- mi impute in STATA

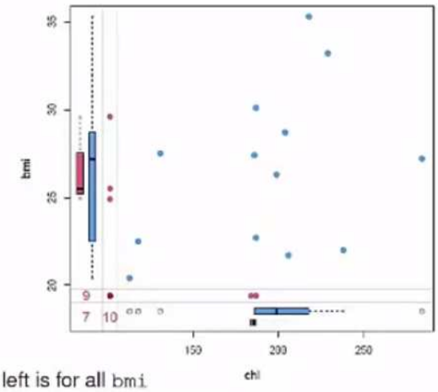
Mice example
• Default method for numeric (continuous) variable is predictive mean matching (pmm)
• Default for 2-level factor is logistic regression (logreg)
• Different method of imputation and set of covariates used can be specified for every variable
• All variables with missing values can be imputed or just a subset of variables
• See van Buuren and Groothuis-Oudshoorn (2011). Mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45, 1-67

• nhanes2 is toy data set (n=25) supplied with mice

• 4 variables: age, body mass index (bmi), hypertension (0-1) hyp, total serum cholesterol (최)

```
require(mice)
head(nhanes2)
    age  bmi  hyp chl
1 20-39   NA <NA>  NA
2 40-59 22.7   no 187
3 20-39   NA   no 187
4 60-99   NA <NA>  NA
5 20-39 20.4   no 113
6 60-99   NA <NA> 184
```

```
require(VIM)
marginplot(nhanes2[, c("chl", "bmi")], col = mdc(1:2), cex = 1.2,
     cex.lab = 1.2, cex.numbers = 1.3, pch = 19)
```



- Blue boxplot on left is for all bmi
- Red boxplot on left is for bmi with missing chl
- If missingness were MCAR, these boxplots would be same

```
nhanes2.imp <- mice(nhanes2, seed = 23109)
```

`summary(nhanes2.imp)` prints info on number of MIs, imputation methods for each variable, covariates used to impute each variable

`complete(nhanes2.imp, action=k)` retrieves the k^{th} completed dataset

```
summary(nhanes2.imp)
Multiply imputed data set
Call:
mice(data = nhanes2, seed = 23109)
Number of multiple imputations:  5
Missing cells per column:
age bmi hyp chl
  0   9   8  10
Imputation methods:
     age     bmi      hyp     chl
      ""   "pmm" "logreg"   "pmm"
VisitSequence:
bmi hyp chl
  2   3   4
PredictorMatrix:
    age bmi hyp chl
age   0   0   0   0
bmi   1   0   1   1
hyp   1   1   0   1
chl   1   1   1   0
Random generator seed value:  23109
```

Regression Example including imputed data

```
fit <- with(nhanes2.imp, lm(chl ~ age + bmi))
round(summary(pool(fit)), 2)
```

	est	se	t	df	Pr(>|t|)	nmis	fmi	lambda
(Intercept)	15.51	57.49	0.27	11.12	0.79	NA	0.38	0.27
age2	44.43	19.02	2.34	9.92	0.04	NA	0.42	0.32
age3	64.94	21.73	2.99	7.62	0.02	NA	0.52	0.41
bmi	5.50	1.95	2.83	11.37	0.02	9	0.37	0.26

`fmi` is "fraction of missing information"
`lambda` is proportion of total variance attributable to imputations

Summary

• Module 1 – general steps in weighting
 - quantiles to estimate
 - goals of estimation and statistical interpretation
 - use of weights to reduce bias and variance
 - effects of weighting on SEs
• Module 2 – specific steps
 - base weights
 - NR adjustments using propensities of response or tree algorithms
 - calibration to external controls
• Module 3 – implementing the steps
 - software for computing weights
• Module 4 – imputing for missing items
 - reasons for imputation
 - methods
 - multiple imputation example using mice