1/26/2010 36-402/608 ADA-II H. Seltman

Handout #5: Missing Data

1. Context
   (a) Single outcome (today) vs. multiple outcome, especially longitudinal data (future week)
   (b) Reasons for missing data: unknown information, sensitive information, unrelated data collection problems
   (c) Types of missing data (cannot be sure which applies to your data)
      i. Missing Completely At Random MCAR): Chance that a value is missing is unrelated to other values on that subject. No possibility of bias. Analyzing complete cases results in loss of power only.
      ii. Missing At Random (MAR): Chance that a value is missing depends on other measured variables for the subject. E.g., if older subjects and males are less likely to answer questions about time spent watching TV, then a model of age, gender, and TV time as predictors of healthiness can be estimated. Analyzing complete cases results in lower power plus the chance for bias. Maximum likelihood methods that include the covariates upon which missingness depends will be unbiased. Other methods, e.g., Generalized Estimating Equations, are biased.
      iii. Not Missing At Random (NMAR): Unknown variables (equivalent to "errors") control the chance of missingness. All standard methods are biased. Methods that attempt to reduce/eliminate bias require strong (often untenable) assumptions. Howard's pet-peeve: "We solved problem X by making assumption Y."

2. Some R techniques
   (a) `y <- c(NA,5,2,NA); is.na(y)` returns `[TRUE FALSE FALSE TRUE]`
   (b) `dtf[dtf$Age==999,] = NA` sets a missing value code
   (c) `mean(y)` returns NA, while `mean(y, na.rm=TRUE)` returns 3.5
   (d) `na.omit(dtf)` returns only the rows of the data frame that have no missing data.
   (e) `m1=lm(y~X1+X2, dtf); m2=lm(y~X1+X2+X3, dtf); anova(m1,m2)` gives "models were not all fitted to the same size of dataset", but an invalid comparison with no warning with, e.g., `AIC(m1)-AIC(m2)`. A valid comparison would need `m1=lm(y~X1+X2, na.omit(dtf[,c("X1","X2","X3","y")]))`.

3. Assessing missing data:
   (a) Percent missing for each variable
   (b) Distribution of number missing for each case
   (c) Distribution of missing value patterns across cases
   (d) Compare Y among missing data patterns
4. Dumb idea: impute missing values as regression estimates (produces overconfident coefficient estimates)
5. Multiple Imputation: State of the Art
   (a) Rough fill-in idea
      i. Fill in anything reasonable, e.g., column means
      ii. Rotate through columns, using regression to fill in missing data
      iii. Continue until there is little change
   (b) Generate multiple filled-in data sets and analyze each one, e.g., lm(y∼X)
   (c) Combine K (usually 3 to 10) estimates in this way:

$$b_{\text{i,MI}} = \frac{1}{K}\sum_{k=1}^{K}\hat{\beta}_{ik}$$

$$W_i = \frac{1}{K}\sum_{k=1}^{K}\text{SE}_i^2$$

$$B_i = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{\beta}_{ik} - \bar{\beta}_i)^2$$

$$V_{i,\text{MI}} = W_i + \left(1 + \frac{1}{K}\right)B_i$$

$$\text{df}_{\text{i,MI}} = (K-1) + \left(1 + \frac{KW_i}{(K+1)B_i}\right)^2$$

   where i is a particular coefficient, "MI" means multiple imputation value, "V" is total sampling variance, "SE" is standard error. So $\sqrt{V_i}$ is the correct MI standard error of b_i.

6. The "mice" multiple imputation package:

```
Xmi <- mice(X, n=5) # create 5 imputed datasets
rslt <- with(Xmi, lm(y~X1+X2+X3)) # fit a model to each dataset
prslt <- pool(rslt) # combine fit results using MI theory
```

   The "pool" result includes

   (a) qbar: the estimated coefficients

   (b) t: the variance covariance matrix of the coeffieients (standard errors are sqrt of diagonal elements)

   (c) df: the effective degrees of freedom

7. Breakout and Discussion