Missing Data Methods

Preliminary Vocabulary:

1. Outcome – a measurement of interest, or endpoint, which varies among the subjects in your study. You want to know why the outcome is "good" for some subjects and "bad" for others.

2. Predictor(s) – a measurement of interest which may effect the outcome of your study. The point of your analysis is to find out is there is a relationship between your predictor(s) and the outcome.

3. Covariates – measurements which are not of intrinsic interest, but need to be accounted for before you can accurately assess the relationship between the predictor(s) and outcome (i.e., patient age).

Warning: There are many types of "missing data" (or what seem like missing data) that can occur in a study. Examples that we will not cover in detail today include:

1. Censored Data: For each subject, you are measuring the amount of time until an event of interest occurs. However, when you end the study, the event will not have occurred yet for some of the subjects.

   The outcome is not really missing, more accurately it is truncated (or, technically, right-censored) for those subjects. The time of truncation is used, as-is, in statistical analyses such as Kaplan-Meier curves, log-rank tests, and Cox regression.

2. Non-Response in Surveys: You attempt to collect survey data and some people respond, but others do not. You have absolutely no outcome, predictor, or covariate information on the non-responders.

   Generally, the best that can be done is to have subjects who have responded "stand-in" for subjects who have not, through "weighting".

   For example, if you sent the survey to 100 Hispanic subjects in Florida and only 50 responded, then each of the 50 responders is counted double (i.e., a weight of 2) in all of your analyses.

3. Missing Repeated Measures (Longitudinal) Data: For each subject, you collect outcome data at multiple points in time (say, baseline, 6 months, 1 year and 2 years). There are two problems that can occur, with slightly different statistical implications.

a. A subject has a gap in their outcome data. For example, they have outcome data at baseline, 1 year and 2 years, but not at 6 months. The fact that they didn't come in at 6 months may be random, or it may be reflective of their condition (i.e., they were too sick to make the appointment). There are a number of ways to try to fill in the gap, or not:
   - Analyze only the data that were collected
   - "Last observation carried forward (LOCF)", fill in the gap at 6 months with the baseline measurement
   - Use a "worst-case" value
   - "Impute" or interpolate the 6 month data based on the measurements at baseline, 1 year and (?) 2 years.

b. A subject drops out of the study. For example, a subject has outcome measurements at baseline and 6 months, but not at 1 or 2 years. A concern here is that the subject may have died and we need to consider whether we want to fill in the missing data or not. If we do, then we can use methods like those above.

4. Missing Data for a Single Outcome or Predictor:
   Each subject in the study has one measurement of outcome and one measurement of the primary predictor. However, some subjects may be missing the outcome or the predictor.

   My suggestion: Remove the subjects with missing data from the database and analyze only subjects that have both the outcome and predictor measured.

   Because of the drop in sample size, you will have less power and less generalizability. This is the price you pay for not getting the requisite data.

   Alternatively: Impute (statistically estimate) the missing data and analyze everyone. Researchers often distrust this method of "making up" data. More on this later.

What We Will Cover Today

Missing Covariate Information: The following suggestions are relevant to missing data in (non-critical??) covariates. If you do not do something about the missing data, you may hurt your power to look at the important predictors. You may also be left with a bias in the coefficients of the important predictors through improper adjustment for confounding and/or reduction of the dataset to a biased subsample.

Approach 1: Let The Sample Size Float (Available Case Analysis)

Use whatever data are available for the current analysis.

Advantages:

- you will have the maximum power for each analysis
- you will have the minimum bias (most generalizability) for each analysis
- you are analyzing “real” data

Disadvantages:

- in going from univariate to multivariable analysis, your results may change for many reasons: confounding; change in sample size/power; change in the cohort being analyzed
- it would be very difficult to characterize exactly which patients are represented by your analyses

Approach 2: Complete Case Analysis

Use only patients who have all of the required covariates

Advantages:

- your sample size/power/generalizability will remain constant
- you can compare the patients in your analyses to those excluded to try to quantify the limits of your generalizability
- you are analyzing “real” data
- you results are unbiased for the cohort being analyzed

Disadvantages:

- your power is minimal
- your generalizability is minimal
- you may be throwing patients out of the analyses because of missing data on covariates which do not get used in the final model

Approach 3: Dummy-Code Missing Data

Create a new category to represent missing responses (i.e., race=1(white); 2(black); 3(hispanic); 4(missing)). Be sure that this dummy variable is forced into all analyses which use any of the original variable's categories. Do not try to interpret the coefficient or significance of the missing data indicator.

Advantages:

- the sample size/power/generalizability for the other covariates is at the maximum possible
- the sample size/power/generalizability for the covariate with missing data is at the maximum for the actually collected "real" data
- only "real" data are used in analyses
- the sample size is constant throughout

Disadvantages:

- the effect estimates for the other covariates may be biased because you have not corrected fully for the confounding effects of the covariate with missing data
- the bias mentioned above will be the worst if the reason that data are missing is related to those other covariates
- the effect estimate for the covariate with missing data could be biased if the reason that data are missing is related to the outcome and to the predictor

Approach 4: Simple Imputation

Replace the missing values with predictions based on patients without missing data. Simplest example: replace missing values with the observed mean/median for that measurement. More complicated: run a regression using the other covariates to predict the missing value.

Advantages:

- sample size is constant and maximal
- if the predictions are accurate, then there is less bias in the coefficients

Disadvantages:

- you are no longer analyzing "real" data
- it feels like circular logic
- the success of the method depends on whether good predictors of the missing variable are available and on the reason for the missing data
- the standard error for the coefficient for the predictor with missing data is underestimated

Approach 5: Multiple Imputation and Extensions

Use a more sophisticated/complicated approach to imputation that gets the standard errors correct and is less likely to produce bias in the effect estimates. Unfortunately, these methods are more computer intensive and often depend on stronger assumptions about the distributions of the covariates and the mechanisms that led to the missing data.

Multiple Imputation

Situation: One or more patients in the database have missing measurements.

Assumption 1: The missingness is not “nonignorable”.

Defintion: Missing data are nonignorable if the probability of missing data depends on the (unobserved) value of the missing data, even after controlling for other variables in the analysis.

Example: If high income households are less likely to report their income, even after adjusting for other variables, then the missing income data are nonignorable.

Assumption 2: All of the measurements are Normally distributed

Process:

1. We assemble a database of all measurements that are related to each other.
2. For a patient who is missing a particular measurement (i.e., age), we look at related information from two sources:
    a. Other covariates that the patient does have recorded, that are correlated with the missing measurement (i.e., if men are generally older in this database than women, and this patient is a man, then the missing age is probably an older age rather than a younger age).
    b. The measurement of interest, as it appears in the other patients who don’t have missing data (i.e., if other patients are around 50 years old, then the missing age is probably around 50 as well).

Example:

	Weight	Height
Patient #1	140	64
Patient #2	120	63
Patient #3	.	65
Patient #4	145	65
Patient #5	130	63
Patient #6	160	66
Patient #7	175	69
Patient #8	170	.
Patient #9	180	72
Patient #10	210	71

Process: Step 3

- To try to impute the missing weight for Patient #3, use the weights of the other 9 patients to get a sense of the usual values in the cohort: Mean Weight = 159 (sd=28)
- Also use the fact that weights are related to heights and we know that Patient #3's height 65 inches:
  Corr(Weight, Height)=.93

Solution: Use Normal-based regression

$$\text{Weight} = -355.9 + 7.705*\text{Height} + \varepsilon$$

where ε is a Normal random variable with variance 144.03

Now, generate a (random) weight for Patient #3 using the regression formula above.

Also

- To try to impute the missing height for Patient #8, use the heights of the other 9 patients to get a sense of the usual values in the cohort: Mean Height = 66.4 (sd=3.4)
- Also use the fact that weights are related to heights and we know that Patient #8's weight 170 pounds:
  Corr(Weight, Height)=.93

Solution: Use Normal-based regression

Height = 49.03 + .112*Weight + ε

where ε is a Normal random variable with variance 2.088

Now, generate a (random) height for Patient #8 using the regression formula above.

New Imputed Data:

	Weight	Height
Patient #1	140	64
Patient #2	120	63
Patient #3	168.74	65
Patient #4	145	65
Patient #5	130	63
Patient #6	160	66
Patient #7	175	69
Patient #8	170	67.52
Patient #9	180	72
Patient #10	210	71

Process: Step 4

With the imputed data inserted into the database, the means and correlation will change:

Mean Weight = 159.9
Mean Height = 66.6
Corr(Weight, Height) = .89

So, you will need to re-run the regressions, get new equations, and generate new (random) values for the missing weight and height.

Process: Step 5

Keep repeating Step 4 until the means and correlation stop changing. Generate one final set of (random) values for the missing weight and height.

→ You now have a dataset with no missing data. Run the analysis you originally intended, for example relating weight to health outcome. The relationship will be captured through a beta coefficient: β and its standard error: $se(\beta)$

Problem: Your analysis treated the imputed data the same as real data. It did not reflect the fact that the imputed data may not be accurate.

Solution: Repeat the 5-Step Imputation process about 5 (no more than 10) times. You will have created 5 complete datasets.

Analyze each of the 5 complete datasets. This will give you 5 β's and 5 se(β)'s.

Your final estimate of the relationship between weight and health outcome will be the average of the 5 β's.

Your final estimate of the variance of β will be

$$\text{Average}(se(\beta)) + [(m+1)/m(m-1)]*\sum(\beta - \text{avg}(\beta))^2$$

Within-imputation variance

Between-imputation variance

Appendix

Methods for Handling Missing CD4 Counts

1. Dummy-Coded: Create two new variables

$$\text{newcd4} = \begin{cases} \text{CD4} & \text{if actual CD4 data are present} \\ 0 & \text{if actual CD4 is missing} \end{cases}$$

$$\text{misscd4} = \begin{cases} 0 & \text{if actual CD4 data are present} \\ 1 & \text{if actual CD4 is missing} \end{cases}$$

General Model:

Mortality = $\alpha + \beta_1$ newcd4 + β_2 misscd4 + other covariates

Model for Patients with Actual CD4 Counts:

Mortality = $\alpha + \beta_1$ CD4 + other covariates

Model for Patients with Missing CD4 Counts:

Mortality = $\alpha + \beta_2$ + other covariates

2. Simple Imputation: If a CD4 count is missing, fill it in with the average CD4 count (among patients who have CD4 data).

First:

Among 49 patients with CD4 data, average CD4 Z-score = -2.69

Second: Create the new predictor:

$$\text{meancd4} = \begin{cases} \text{CD4} & \text{if actual CD4 data are present} \\ -2.69 & \text{if actual CD4 is missing} \end{cases}$$

3. Regression Imputation: If a CD4 count is missing, fill it in with the predicted CD4 count based on a regression model using patients who have CD4 data.

First:
Among 49 patients with CD4 data, run the linear regression model:

Predicted CD4 = Intercept -.202 Year of Diagnosis +.177 Sex
- .319 White - 1.31 Transfusion + .227 Age<1
- .064 Encephalopathy - .235 LIP + .535 Wasting
+ .077 IGGZ - .018 IGMZ

(Model R^2 = 58%)

Second: Create the new predictor:

$$\text{impcd4} = \begin{cases} \text{CD4} & \text{if actual CD4 data are present} \\ \text{Predicted CD4} & \text{if actual CD4 is missing} \end{cases}$$

Modeling Results

	Relative Risk and P-Value					
Sample Size / Deaths:	68/43	49/27	49/27	68/43	68/43	68/43
Predictors						
Year of Diagnosis	.83 p=.02	.84 p=.07	.92 p=.37	.85 p=.06	.79 p<.01	.79 p<.01
Encephalopathy	2.3 p=.02	6.6 p<.01	3.7 p<.01	4.4 p<.01	3.5 p<.01	3.0 p<.01
Wasting	4.8 p=.01	6.7 p<.01	7.2 p<.01	6.2 p<.01	4.4 p<.01	4.5 p<.01
IGG Z-score	.87 p<.01	.88 p=.08	.86 p=.02	.80 p<.01	.88 p=.01	.89 p=.03
CD4 Z-score		.44 p<.01				
newcd4				.47 p<.01		
misscd4				33.8 p<.01		
meancd4					.52 p<.01	
impcd4						.64 p=.01