BST 210
Applied Regression Analysis

"Well, this certainly explains much of the company's missing data. Who else thought the 'DEL' key on their computer was for delegating work?"

Save the Date!

**** December 19: 11:30am-1:00pm FXB G12 ****

Special Program for our final class!

featuring guest speakers and their seminal contributions to the field of statistics and modeling

And of course a celebration!

**** We expect *ALL BST 210 students* to attend. ****

(no missing data, please)

Lecture 27
Plan for Today

Missing Data

- Descriptions of missing data
- Mechanisms:
  - MCAR
  - MAR
  - MNAR (non-ignorable)
- Exploration of missingness in a data set
- Analysis under missingness
  - Naïve approaches (CC)
  - Less naïve approaches (Multiple Imputation methods)

Missing Data

- Missing data is a *common occurrence* in the analysis of data, not just longitudinal or cohort outcomes, but even for cross-sectional data (lab test not done, questionnaire not filled out…)

- So, what's the dilemma?

  … *`Once randomized, always analyzed'…*

  and ignoring missing data when we analyze can introduce not only lack of efficiency, but also bias in our resulting inference.

- Best advice: recognize and account for missingness!

Missing Data

- Missing data could involve either (or both)

  (1) missing covariates or

  (2) missing outcomes

- Missing data (when not accounted for) may lead to

  - *Loss of* efficiency: there is often some loss of information

  - *bias*: invalid inference may result

Missing Data

- Unfortunately, *given data are missing*, we cannot empirically verify any particular missing data assumption

- Thus, exploratory data analysis and knowledge of context and subject matter should guide us as to potential plausible missing data assumptions

- We can then approach the data analysis stage with some confidence, but can also report analyses under various assumptions (sensitivity analysis)

Reasons for Missing Data

-
-
-
-

Reporting of Missing Data

- Sample sizes and missingness should be clear in the reporting of data
- For categorical values, report values and denominators (and %)
- For continuous variables, report sample sizes if appropriate
- Every regression model run should have a clear sample size attached to it
- Sample sizes could appear in table titles, column headings, or footnotes
- “Missing < 5% of values except for …” is commonly seen

Caveats re Reporting of Missing Data

- In many studies with missing data or dropouts, one would compare the baseline characteristics of subjects with data vs. subjects without data

- If the proportion of subjects with missing data was small and the missingness looked unrelated to baseline characteristics, any bias due to missing data may be small in general

- But you can never be 100% sure, given the missingness could be related to the outcome that is missing

Missing Data Mechanisms

- Don Rubin developed a classification system for missing data mechanisms – see the seminal (1976) paper here

- We introduce these classifications in this simple setting:
  - Assume that we have a fully observed covariate (x) and a partially observed outcome (Y).
  - Thus, Y might be missing for some observations.
  - Let R denote an indicator for missingness:

    $R = 1$ when Y is missing,

    $R = 0$ when Y is observed
  - We are then interested in $P(R = 1 \mid x, Y)$

Missing Completely at Random (MCAR)

- The missing values in *Y* are said to be missing completely at random (MCAR) if the missingness is independent of both *Y* and *x*

- Thus, those subjects with a missing outcome *Y* do not differ systematically (in terms of *Y* or *x*) to those with *Y* observed

- That is,

$$P(R = 1 \mid x, Y) = P(R = 1)$$

- Is this a restrictive assumption?

Examples of Data Missing Completely at Random

-
-
-
-

Checking for MCAR?

- Given the observed data, we *could* assess whether the covariate *x* is associated with the missingness indicator *R*
- If it is, we can conclude the data are not MCAR
- However, if it is not, we cannot necessarily conclude that the data are MCAR, because the missingness *R* could still depend on the outcomes *Y*, but we just don't have all values of *Y* to assess this
- Bottom line with MCAR is that the analysis of complete cases (ignoring outcomes with missing values *Y*) can be performed without any concern about bias, just concern about loss of efficiency

Missing at Random (MAR)

- The missing values in *Y* are said to be missing at random (MAR) given the covariate *x,* if the missingness is independent of *Y* given *x*

- Thus, among those subjects with the same value of the covariate *x*, missingness in the outcome *Y* is independent of the value of *Y*

- Thus,

$$P(R = 1 \mid x, Y) = P(R = 1 \mid x)$$

- This is still a very restrictive assumption, and impossible to assess

Examples of Data Missing at Random

-
-
-
-

Checking for MAR?

- Given the observed data, we *cannot* assess whether the missingness indicator *R* is independent of *Y* given the values of the covariate *x*

- To do this, we would need to check, for each value of *x*, that those with missing *Y* had similar or different outcome values than those with observed *Y* – *but we don't have these missing Y!*

MAR

- It is important to note that MAR implies that

  $$f(Y \mid x, R = 0) = f(Y \mid x, R = 1)$$

- That is, the distribution of the outcome variable Y given covariate x, is the same when we have a missing observation Y or a non-missing observation Y

- MLEs? Bottom line with MAR is that the analysis of complete cases (ignoring outcomes with missing values *Y*) can be performed without any concern about bias if maximum likelihood methods are used, though we still have a concern about loss of efficiency

- Thus, we can still estimate the betas, etc based on the values we've observed (observed likelihood), and don't need the values we haven't observed (since they are distributed the same as the observed)

- (Note: Use of other methods however, such as generalized estimating equations, could lead to bias)

Missing Not at Random (MNAR)

- The missing values in *Y* are said to be missing not at random (MNAR) if they are *neither* MCAR nor MAR

- Effectively in this case, the chance of observing *Y* depends on the value of *Y*, even after conditioning on the value of the covariate *x*

- A consequence is that

$$f(Y \mid x, R = 0) \neq f(Y \mid x, R = 1)$$

…since the probability of missingness here depends on the very outcome Y! (It may also depend on x, but given known values of x, the issue here is that the missingness also depends on Y)

Examples of Data Missing Not at Random

-
-
-
-

Checking for MNAR?

- Given the observed data, we *cannot* assess whether the MNAR is occurring or not, i.e., the data cannot tell us how the missing *Y* values differ from observed *Y* values, given the value of the covariate *x*

- Bottom line with MNAR is that particular assumptions have to be made in order to perform analyses, and sensitivity analyses should be performed to see how inferences vary with different assumptions, i.e., how biases might arise

- Likelihood based approaches without consideration of MNAR will cause not only lack of efficacy, but also potential (and likely) bias in our inferences

Analysis Under Missingness

- Naïve Approaches
  - Complete Case Analysis
  - Assume the “worst outcome” for missing responses
  - Assume “no change from baseline” for missing responses
  - Make a “missing category” for a categorical variable and include that in the modeling
  - Fill in the “mean value” for a missing continuous variable
  - there are many more!

- Less Naïve Approaches
  - Multiple Imputation
  - Inverse proportional weighting
  - there are many more! (selection models, pattern mixture models, …)

- Note that in most clinical epidemiological studies we tend to see more MAR data than MCAR or MNAR.

Complete Case (CC) Analysis

- The complete case (CC) approach to analysis under missing data involves using data only from those subjects for whom all of the variables involved with our analysis (here both *x* and *Y*) are observed (what we've done most of this course!)

- This is the default approach of most statistical packages when performing analyses *(know what your software is doing with missing data or you could inadvertently introduce bias!)*

- Estimation may be biased if the complete cases (CC) differ systematically from the incomplete cases

Complete Case (CC) Analysis

- CC analyses are valid provided the probability of being a CC is independent of outcome, given the covariates in the model of interest

- CC analyses are unbiased if the missingness is independent of the outcome *Y*, i.e.,

$$P(R = 1 \mid Y) = P(R = 1)$$

- CC analyses are also unbiased if we condition on the cause of the missingness in the analysis, e.g., if the missingness is dependent on the covariate *x* and we adjust for *x* in the analysis

Multiple Imputation (MI)

- Impute or “fill in” each missing value multiple times, resulting in multiple completed data sets
- Analyze each completed data set individually
- Combine the estimates appropriately to get overall estimates and appropriate measures of variability
- By using the data from all cases now, estimates based on MI are generally more efficient than estimates from CC, and may even remove some bias in CC estimates
- But, the results depend on a correct model for imputation

Multiple Imputation

- The overall estimates for our parameters are the average of the estimates from the completed data sets

- The overall variance estimates are based on combining the average of the variance estimates ("within imputation variance") and the uncertainty due to the missing data ("between imputation variance")

- Number of imputations 'm' is often on the order of 5 or 10

Example: HERS Study (Vittinghoff text)

- The Heart and Estrogen/Progestin Study (HERS) was a clinical trial of hormone therapy for the prevention of recurrent heart attacks and death among 2,763 post-menopausal women with existing coronary heart disease

- Here we look at a few predictors of systolic blood pressure, some of which may be missing themselves (especially glucose)

Example: HERS Study (Vittinghoff text)

- The full data set considering a systolic blood pressure outcome variable has 1,871 observations
- A CC analysis of a linear regression model predicting systolic blood pressure as a function of glucose, race (white vs. non-white), and BMI uses 1,385 observations (74% of the observations)
- Much of the reduced sample size is due to missing glucose levels. Using an MI approach to impute glucose results in 1,750 observations being used (94% of the observations)
- There is some change in the MI beta estimates (especially BMI), and s.e. estimates get smaller (compared to CC)

Multiple Imputation

- MI gives unbiased (or asymptotically unbiased) estimates provided the data are MAR and the imputation model is correctly specified

- The plausibility of the MAR assumption can be guided by data analysis and subject matter knowledge

- We often include more variables in the imputation model than in the main regression model of interest, as including these extra variables in the imputation model may increase the likelihood of MAR holding

Multiple Imputation

- Main point #1: We still need to have correctly specified the model(s) that are used to impute missing values

- Main point #2: If the missingness is not at random, you may need a sensitivity analysis under varying plausible assumptions that you make up

- Main point #3: We should always do analyses both with and without imputation, with full reporting of results and approaches, and keep large picture with both pros and cons of various aspects of our entire analysis process, in mind

Other MI Approaches

- Imputation of missing categorical variables using logistic, multinomial, or ordinal logistic regression
- Imputation of various patterns of missing variables using multivariate normal assumptions
- Lot's more!

Inverse Proportional Weighting

- An alternative approach is to perform a CC analysis, but weight the data used by the inverse of the probability of having the data be observed
- Those subjects who had a small chance of being observed are given higher weights, to compensate for similar subjects with missing data
- Thus, we need to model how missingness depends on fully observed variables

Practice Example

- For each example below for the outcome BMI and the covariate age, describe the type of missingness (MCAR, MAR, MNAR) and how to adjust for this.

1. BMI is missing for some subjects because only one scale was available. Some subjects didn't have their weight measured if the scale was occupied.

2. BMI is missing for younger subjects because they had to go to school and couldn't wait to have their BMI measured.

3. BMI is missing for older subjects with a higher BMI because they didn't want their weight measured.

Coming Up

… Let's next review all of this and do a multiple imputation example in R (in coordination with the LabWeek15) →

Also don't forget:

- Thursday review for Final Exam
- Save the date! ** Thursday December 19: 11:30am-1:00pm **

Special Program for our final class!
featuring guest speakers and their seminal contributions to the field of statistics and modeling