Missing Data
Analysis and Imputation

Introduction

- Missing data arise in almost all serious statistical analyses.
- In classical analysis such as multiple regression, majority of statistical packages automatically excludes all cases in which any of the inputs are missing.
- This approach is called a complete-case analysis.
- This can limit the amount of information available in the analysis, especially if the model includes many inputs with potential missingness.
- Things become more difficult when predictors have missing values.

Missingness Mechanism

- To decide how to handle missing data, it is helpful to know why they are missing.
- We consider four general “missingness mechanisms,” moving from the simplest to the most general.

Missingness Mechanism

- ***Missingness Completely At Random (MCAR)***
  - Probability of missingness is the same for all units;
  - Throwing out cases with missing data does not bias your inferences.
- ***Missingness At Random (MAR)***
  - Completely at random for each categories of another variable;
  - Probability a variable is missing depends only on available information.
- ***Missingness that depends on unobserved predictors***
  - If missingness depends on information that has not been recorded and also predicts the missing values.
- ***Missingness that depends on the missing value itself***
  - probability of missingness depends on the (potentially missing) variable itself

How to Prove MAR – Impossible Maybe!

- **MAR is relatively easy to handle.**
  - Simply include as regression inputs all variables that affect the probability of missingness.
  - But how to know if data really are MAR, or whether the missingness depends on unobserved predictors or the missing data themselves (unobserved censoring indicator variable).
  - The fundamental difficulty is that unobserved variables are unobserved!!

Missing-data Analysis Methods that Discard Data

- Many missing data approaches simplify the problem by throwing away data
- This may lead to
  - Biased estimates,
  - Estimates with large standard error.

Missing-data Analysis Methods that Discard Data – Cont.

- ***Complete-case Analysis***
  - A direct approach to missing data is to exclude them.
  - In the regression context, this is usually called complete-case analysis, that means excluding all units for which the outcome or any of the inputs are missing.
  - Two problems arise with complete-case analysis:
    - If the units with missing values differ systematically from the 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, so that most of the data would be discarded for the sake of a simple analysis.

Missing-data Analysis Methods that Discard Data – Cont.

- ***Available-case Analysis***
  - To analyse each variable based on the available data for that particular variable.
    - This approach has the problem that different analyses will be based on different subsets of the data and thus will not necessarily be consistent with each other.
    - In addition, if the non-respondents differ systematically from the respondents, this will bias the available-case summaries.
- ***Complete-variables analysis***
  - This is when a researcher simply excludes a variable or set of variables from the analysis because of their missing-data rates.
    - This may lead to omission of a variable that is necessary to satisfy the assumptions required for interpretations.

Missing-data Analysis Methods that Discard Data – Cont.

- ***Nonresponse Weighting***
  - Easy to use when only one - mainly categorical - variable has missing data.
    - We could build a model to predict the nonresponse in that variable using all the other variables.
    - The inverse of predicted probabilities of response from this model could then be used as survey weights to make the complete-case sample representative (along the dimensions measured by the other predictors) of the full sample.
    - There is the potential that standard errors will become erratic if predicted probabilities are close to 0 or 1.
    - Also, this method becomes more complicated when there is more than one variable with missing data.

Missing-data Analysis Methods that Retain Data

- Rather than removing variables or observations with missing data, another approach is to fill in or "impute" missing values.
- A variety of imputation approaches can be used that range from extremely simple to rather complex.
- These methods keep the full sample size, which can be advantageous for bias and precision; however, they can yield different kinds of bias.
- Whenever a single imputation strategy is used, the standard errors of estimates tend to be too low.
- The intuition here is that we have substantial uncertainty about the missing values, but by choosing a single imputation we in essence pretend that we know the true value with certainty.

Missing-data Analysis Methods that Retain Data – Cont.

- ***Mean Imputation***
  - The easiest way to impute is to replace each missing value with the mean of the observed values for that variable.
    - This can severely distort the distribution for this variable, leading to complications with summary measures.
    - Underestimates of the standard deviation.
    - Moreover, mean imputation distorts relationships between variables by "pulling" estimates of the correlation toward zero.

Missing-data Analysis Methods that Retain Data – Cont.

- ***Last Value Carried Forward***
  - The easiest way to impute is to replace each missing value with the mean of the observed values for that variable.
    - In evaluations of interventions (where pre-treatment measures of the outcome variable are also recorded), one missing values imputation could be to replace missing outcome values with the pre-treatment measure.
    - This is often thought to be a conservative approach as it would lead to underestimates of the true treatment effect.
    - There are situations in which this strategy can be anti-conservative, for instance in the case of case-control studies.

Missing-data Analysis Methods that Retain Data – Cont.

- ***Using Information From Related Observations***
- Suppose we are missing data regarding the income of fathers of children in a dataset. Why not fill these values in with mother's report of the values?
  - This is a plausible strategy, although these imputations may propagate measurement error.
  - Also we must consider whether there is any incentive for the reporting person to misrepresent the measurement.

Missing-data Analysis Methods that Retain Data – Cont.

- ***Indicator variables for missingness of categorical predictors***
- For unordered categorical predictors, a simple and often useful approach to imputation is to add an extra category for the variable indicating missingness.

Missing-data Analysis Methods that Retain Data – Cont.

- ***Indicator variables for missingness of continuous predictors***
- A popular approach in the social sciences is to include for each continuous predictor variable with missingness an extra indicator identifying which observations on that variable have missing data.
- Then the missing values in the partially observed predictor are replaced by zeroes or by the mean.
- This strategy is prone to yield biased coefficient estimates for the other predictors included in the model.

Missing-data Analysis Methods that Retain Data – Cont.

- ***Imputation based on logical rules***
- Logical relationship between the variable with missing values to another variable. As an example, income might be missing, but if the number of hours of work per week is zero, then income could be assumed zero.

Random Imputation of a Single Variable

- When more than a trivial fraction of data are missing, we prefer to perform imputations more formally.
- ***Simple random imputation***
  - The simplest approach is to impute missing values of earnings based on the observed data for this variable. Does it make sense?

R code

```
random.imp <- function (a){
  missing <- is.na(a)
  n.missing <- sum(missing)
  a.obs <- a[!missing]
  imputed <- a
  imputed[missing] <- sample (a.obs, n.missing, replace=TRUE)
  return (imputed)
}
 earnings.imp <- random.imp (earnings)
```

Random Imputation of a Single Variable – Cont.

- ***Zero Coding and Top-coding***
  - We shall fit the regression model to those respondents whose earnings were observed and positive.
  - In addition, we shall top-code all earnings at $100,000—that is, all responses above this value will be set to $100,000. The top-coding reduces the sensitivity of the results to the highest values.
  - By top-coding we'll lose information but as main use of earnings in this survey is to categorize families into income quantiles, for which purpose top-coding at $100,000 has no effect.

R code

```
topcode <- function (a, top){
  return (ifelse (a>top, top, a))
}
earnings.top <- topcode (earnings, 100)   # earnings are in $thousands
hist (earnings.top[earnings>0])
```

Random Imputation of a Single Variable – Cont.

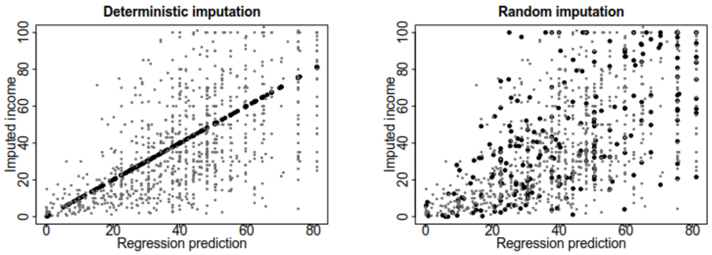
- ***Using regression predictions to perform deterministic imputation***
  - A simple and general imputation procedure that uses individual-level information.
  - We begin by setting up a data frame with all the variables we shall use in our analysis, and then fit a regression to positive values of earnings.

```
sis <- data.frame (cbind (earnings, earnings.top, male, over65, white,
  immig, educ_r, workmos, workhrs.top, any.ssi, any.welfare, any.charity))

lm.imp.1 <- lm (earnings ~ male + over65 + white + immig + educ_r +
  workmos + workhrs.top + any.ssi + any.welfare + any.charity,
  data=SIS, subset=earnings>0)
```

Random Imputation of a Single Variable – Cont.

- ***Using regression predictions to perform random imputation***
  - We can put the uncertainty back into the imputations by adding the prediction error into the regression.



Random Imputation of a Single Variable – Cont.

- ***Two-stage modelling to impute a variable that can be positive or zero***
- suppose that the workhrs and workmosvariables were not available, so that we could not immediately identify the cases with zero earnings.

We would then impute missing responses to the earnings question in two steps:

  - First, imputing an indicator for whether earnings are positive, and,
  - Second, imputing the continuous positive values of earnings.

Random Imputation of a Single Variable – Cont.

- ***Two-stage modelling to impute a variable that can be positive or zero***
- Mathematically, we would impute earnings y given regression predictors X in a two-step process, defining:

$$y = I^y y^{\text{pos}}$$

- Where $I^y = 1$ if $y > 0$ and 0 otherwise, and $y^{pos} = y$ if $y > 0$. The first model is a logistic regression for I^y

$$\Pr(I_i^y = 1) = \text{logit}^{-1}(X_i\alpha),$$

and the second part is a linear regression for the square root of y^{pos}:

$$\sqrt{y_i^{\text{pos}}} \sim \text{N}(X_i\beta, \sigma^2).$$

Random Imputation of a Single Variable – Cont.

- ***Matching and hot-deck imputation***
- A different way to impute is through matching: for each unit with a missing y, find a unit with similar values of X in the observed data and take its y value.

Random Imputation of Several Variables

- ***Routine multivariate imputation***
- The direct approach to imputing missing data in several variables is to fit a multivariate model to all the variables that have missingness.
- The difficulty of this approach is that it requires a lot of effort to set up a reasonable multivariate regression model.

Random Imputation of Several Variables

- ***Iterative regression imputation***
- A different way to generalize the univariate methods of the previous section is to apply them iteratively to the variables with missingness in the data.
- If the variables with missingness are a matrix Y with columns Y(1), . . . , Y(K) and the fully observed predictors are X, this entails first imputing all the missing Y values using ***Routine multivariate imputation***
- Then imputing Y(1) given Y(2), . . . , Y(K) and X; imputing Y(2) given Y(1), Y(3), . . . , Y(K) and X (using the newly imputed values for Y(1)), and so forth, randomly imputing each variable and looping through until approximate convergence.