Notes: Flexible Imputation of Missing Data – Ch2 Multiple Imputation

Yingbo Li

08/25/2020

Table of Contents

Notations

- m: number of multiple imputations
- Y: data of the sample
  - Includes both covariates and response
  - Dimension $n \times p$
- R: observation indicator matrix, known
  - A $n \times p$ 0-1 matrix
  - $r_{ij} = 0$ for missing and 1 for observed
- Y_{obs}: observed data
- Y_{mis}: missing data
- $Y = (Y_{\text{obs}}, Y_{\text{mis}})$: complete data
- ψ: the parameter for the missing mechanism
- θ: the parameter for the full data Y

Concepts of MCAR, MAR, and MNAR, with notations

- Missing completely at random (MCAR)

$$P(R = 0 \mid Y_{\text{obs}}, Y_{\text{mis}}, \psi) = P(R = 0 \mid \psi)$$

- Missing at random (MAR)

$$P(R = 0 \mid Y_{\text{obs}}, Y_{\text{mis}}, \psi) = P(R = 0 \mid Y_{\text{obs}}, \psi)$$

- Missing not at random (MNAR)

$$P(R = 0 \mid Y_{\text{obs}}, Y_{\text{mis}}, \psi) \text{ does not simplify}$$

Ignorable

- The missing data mechanism is ignorable for likelihood inference (on θ), if
    1. MAR, and
    2. Distinctness: the parameters θ and ψ are independent (from a Bayesian's view)
- If the nonresponse if ignorable, then

$$P(Y_{\text{mis}} \mid Y_{\text{obs}}, R) = P(Y_{\text{mis}} \mid Y_{\text{obs}})$$

Thus, if the missing data model is ignorable, we can model θ just using the observed data

Goal of multiple imputation

- Note: for most multiple imputation practice, this goal is to train a (predictive) model with as small variances of the parameters as possible
- Q: estimand (the parameter to be estimated)
- $\hat{Q}$: estimate
  - Unbias

$$E(\hat{Q} \mid Y) = Q$$

  - Confidence valid:

$$E(U \mid Y) \geq V(\hat{Q} \mid Y)$$

    where U is the estimated covariance matrix of $\hat{Q}$, the expectation is over all possible samples, and $V(\hat{Q} \mid Y)$ is the variance caused by the sampling process

Within-variance and between-variance

$$E(Q \mid Y_{\text{obs}}) = E_{Y_{\text{mis}}|Y_{\text{obs}}}\{E(Q \mid Y_{\text{obs}}, Y_{\text{mis}})\}$$
$$V(Q \mid Y_{\text{obs}}) = \underbrace{E_{Y_{\text{mis}}|Y_{\text{obs}}}\{V(Q \mid Y_{\text{obs}}, Y_{\text{mis}})\}}_{\text{within-variance}} + \underbrace{V_{Y_{\text{mis}}|Y_{\text{obs}}}\{E(Q \mid Y_{\text{obs}}, Y_{\text{mis}})\}}_{\text{between variance}}$$

- Within-variance: average of the repeated complete-data posterior variance of Q, estimated by

$$\bar{U} = \frac{1}{m} \sum_{l=1}^{m} \bar{U}_l,$$

  where $\bar{U}_l$ is the variance of $\hat{Q}_l$ in the lth imputation

- Between-variance: variance between the complete-data posterior means of Q, estimated by

$$B = \frac{1}{m-1} \sum_{l=1}^{m} \left(\hat{Q}_l - \bar{Q}\right)\left(\hat{Q}_l - \bar{Q}\right)', \quad \bar{Q} = \frac{1}{m} \sum_{l=1}^{m} \hat{Q}_l$$

Decomposition of total variation

- Since $\bar{Q}$ is estimated using finite m, the contribution to the variance is about B/m. Thus, the total posterior variance of Q can be decomposed into three parts:

$$T = \bar{U} + B + B/m = \bar{U} + \left(1 + \frac{1}{m}\right) B$$

- $\bar{U}$: the conventional variance, due to sampling rather than getting the entire population.
- B: the extra variance due to missing values
- B/m: the extra simulation variance because $\bar{Q}$ is estimated for finite m
  - Traditionally choices are $m = 3, 5, 10$, but the current advice is to use a larger m, e.g., $m = 50$

Properness of an imputation procedure

- An imputation procedure is confidence proper for complete-data statistics $\hat{Q}, U$, if it satisfies the following three conditions approximately at large m

$$E\left(\bar{Q} \mid Y\right) = \hat{Q}$$
$$E\left(\bar{U} \mid Y\right) = \hat{U}$$
$$\left(1 + \frac{1}{m}\right) E(B \mid Y) \geq V(\bar{Q})$$

  - Here $\hat{Q}$ is the complete-sample estimator of Q, and U is its covariance
  - If we replace the $\geq$ by $>$ in the third formula, then the procedure is said to be proper
  - It is not always easy to check whether a procedure is proper.

Scope of the imputation model

- Broad: one set of imputations to be used for all projects and analyses
- Intermediate: one set of imputations per project and use this for all analyses
- Narrow: a separate imputed dataset is created for each analysis
- Which one is better: depends on the use case

Variance ratios

- Proportion of variation attributable to the missing data

$$\lambda = \frac{B + B/m}{T}$$

  - If $\lambda > 0.5$, then the influence of the imputation model on the final result is larger than that of the complete-data model

- Relative increase in variance due to nonresponse

$$r = \frac{B + B/m}{\bar{U}} = \frac{\lambda}{1 - \lambda}$$

- Fraction of information about Q missing due to nonresponse

$$\gamma = \frac{r + 2/(\nu + 3)}{1 + r} = \frac{\nu + 1}{\nu + 3}\lambda + \frac{2}{\nu + 3}$$

  - Here, ν is the degrees of freedom (see next)
  - When ν is large, γ is very close to λ

Degrees of freedom (df)

- The degrees of freedom is the number of observations after accounting for the number of parameters in the model.
- The "old" formula (as in Rubin 1987): may produce values larger than the sample size in the complete data

$$\nu_{\text{old}} = (m-1)\left(1 + \frac{1}{r^2}\right) = \frac{m-1}{\lambda^2}$$

- Let ν_{com} be the conventional df in a complete-data inference problem. If the number of parameters in the model is k and the sample size is n, then $\nu_{\text{com}} = n - k$. The estimated observed data df that accounts for the missing information is

$$\nu_{\text{obs}} = \frac{\nu_{\text{com}} + 1}{\nu_{\text{com}} + 3}\nu_{\text{com}}(1-\lambda)$$

- Barnard-Rubin correction: the adjusted df to be used for testing in multiple imputation is

$$\nu = \frac{\nu_{\text{old}}\nu_{\text{obs}}}{\nu_{\text{old}} + \nu_{\text{obs}}}$$

A numerical example

```r
## Load the mice package
library(mice);
imp <- mice(nhanes, print = FALSE, m = 10, seed = 24415)
fit <- with(imp, lm(bmi ~ age))
est <- pool(fit); print(est, digits = 2)
```

```
## Class: mipo    m = 10
##          term  m estimate ubar    b   t dfcom   df  riv l
## 1 (Intercept) 10     30.8  3.4 2.52 6.2    23  9.2 0.82
## 2         age 10     -2.3  0.9 0.39 1.3    23 12.3 0.48
```

- Columns `ubar`, `b`, and `t` are the variances
- Column `dfcom` is ν_{com}
- Column `df` is the Barnard-Rubin correction ν

T-test for regression coefficients

- Use the Barnard-Rubin correction of ν as the shape parameter of t-distribution.

```
print(summary(est, conf.int = TRUE), digits = 1)
```

```
##          term estimate std.error statistic df p.value 2.5
## 1 (Intercept)       31         2        12  9   5e-07
## 2         age       -2         1        -2 12   7e-02
```

Imputation evaluation criteria

- The following criteria are useful in simulation studies (when you know the true Q)

1. Raw bias (RB): upper limit 5%

$$\text{RB} = \left| \frac{E\left(\bar{Q}\right) - Q}{Q} \right|$$

2. Coverage rate (CR): A CR below 90% for the nominal 95% interval is bad
3. Average width (AW) of confidence interval
4. Root mean squared error (RMSE): the smaller the better

$$\text{RMSE} = \sqrt{\left(E\left(\bar{Q}\right) - Q\right)^2}$$

Imputation is not prediction

- Shall we evaluate an imputation method by examine how it can closely recover the missing values?
  - For example, using the RMSE to see if the imputed values $\dot{y}_i$ are close to the true (removed) missing data y_i^{mis}?

$$\text{RMSE} = \sqrt{\frac{1}{n_{\text{mis}}} \sum_{i=1}^{n_{\text{mis}}} \left(y_i^{\text{mis}} - \dot{y}_i\right)^2}$$

- NO! This will favor least squares estimates, and it will find the same values over and over; and thus it is single imputation. This ignores the inherent uncertainty of the missing values.

When not to use multiple imputation

- For predictive modeling, if the missing values are in the target variable Y, then complete-case analysis and multiple imputation are equivalent.
- Two special cases where listwise deletion is better than multiple imputation

1. If the probability to be missing does not depend on Y
2. If the complete data model is logistic regression, and the missing data are confined to Y, not X

References

- Van Buuren, S. (2018). Flexible Imputation of Missing Data, 2nd Edition. CRC press.
  - https://stefvanbuuren.name/fimd/
- Rubin, D. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons.