Diagnosing Multiple Imputation Models Using the Propensity Score

Yulei He, Guangyu Zhang, Erin Dienes,
Nat Schenker
National Center for Health Statistics

National Conference on Health Statistics
2015

Outline

- Background
- Problem
- Main Results
- Discussion

Background

- Multiple imputation (MI) is a popular approach to missing data problems
- A lack of principled diagnostic procedures for imputations
- Practitioners tend to compare the distribution of observed with imputed data
- These ad-hoc comparisons can be based on erroneous assumptions and reach misleading conclusions

Goal and Motivation

- Comparing the distribution of observed and imputed data is a natural diagnostic strategy
- The key is to assess the balance between the imputed and observed data
- The balancing of covariates also necessitates valid causal inference in observational studies, assisted by the propensity score
- Aim to establish some principled diagnostic procedures for imputation, borrowing the idea of propensity score analysis in observational studies

Set-up and Notations

- Consider a single incomplete variable Y and fully observed covariates X
- X can be multidimensional
- Y_{mis} and Y_{obs} are the missing and observed case of Y
- R is the response indicator, R=1 if Y is observed and 0 otherwise
- The missingness of Y is strongly ignorable
- Define the response propensity score is $Pr(R=1|X)=g(X)$, $0<g(X)<1$
- Y_{imp} denote the imputations for Y_{mis}

Main Logic

- Establish some balancing relationships between Y_{mis} and Y_{obs} in the considered scenario
- If Y_{mis} is correctly imputed by Y_{imp} ,then we would expect Y_{imp} also satisfies these balancing relationships
- Considerable violations of the balancing relationships between Y_{imp} and Y_{obs} would suggest some inadequacy of the former

Balancing Properties

- From a Bayesian perspective, suppose Y_{imp} is drawn from a correctly specified imputation model, then as the sample size increases with a fixed proportion of missingness,

  $$\Pr(Y_{imp} \mid g(X),R=0) \rightarrow \Pr(Y_{obs} \mid g(X),R=1) \quad (5)$$

  $$\Pr(X \mid Y_{imp}, g(X),R=0) \rightarrow \Pr(X \mid Y_{obs}, g(X),R=1) \quad (6)$$

- The lack of strict equivalence in Eqs. (5) and (6) (as opposed to Eqs. (3) and (4)) is due to the uncertainty of the estimation in the imputation model parameter

Comparing the Imputed with Observed Data

- Eq. (5) suggests that
- Compare $\Pr(Y_{imp} | g(X), R=0)$ and $\Pr(Y_{obs} | g(X), R=1)$, the conditional distribution of the outcome Y on g(X)
- Eq. (6) suggests that
- Compare $\Pr(X|Y_{imp}, g(X), R=0)$ and $\Pr(X|Y_{obs}, g(X), R=1)$, the conditional distribution of the covariates X on Y and g(X)
- A considerable lack of similarity in either case would suggest some inadequacy of the imputations
- For the 2nd comparison, we can consider some meaningful scalar functions f(X) for comparison
- f(X) could be each of the covariates or interactions among them

Practical Implementations

- Comparing two conditional distributions is not a simple task
- Do not take an hypothesis testing approach to this problem
- In observational studies, diagnostic procedures have been developed to assess whether the estimated propensity score balances the distribution of covariates
- In our cases, suppose g(X) is correctly estimated, then it should balance between the imputed and observed values, as implied by Eqs. (5) and (6)
- Take advantage of these diagnostics procedures in our setting

Comparison through Matching

- Usually the missingness proportion is less than 50%
- Estimate the propensity score g(X) and check its adequacy by assessing the balance of covariates
- Compare $\Pr(Y_{imp} | g(X), R=0)$ and $\Pr(Y_{obs} | g(X), R=1)$: construct a one-to-one matched sample between missing and observed cases using g(X), and compare Y_{imp} and Y_{obs} on the matched sample
- Compare $\Pr(X|Y_{imp}, g(X), R=0)$ and $\Pr(X|Y_{obs}, g(X), R=1)$: construct a one-to-one matched sample between missing and observed cases using both Y and g(X), and compare the respective f(X) on the matched sample

Balancing Diagnostic Statistics

- Many diagnostics statistics are available in observational studies
- We focus on two
- The standardized difference between the matched sample (STDDIFF)
- The variance ratio between the matched sample (VARRATIO)
- The evidence of balance is strong if STDDIFF is close to 0 and VARRATIO is close to 1
- Common criteria: e.g., balance achieved if STDDIFF < 10%
- In our context, we average these diagnostics over multiply imputed datasets
- We calculate the frequency (probability) that these diagnostics exceed some thresholds

Numerical Example

- A 10% random subset of 2002 US Natality public-use data (sample size around 40K)
- The incomplete variable Y is gestational age (DGESTAT), the rate of missing is around 18%
- Covariates X include a wide variety of demographic and health characteristics
- The covariate birthweight (DBIRWT) has the largest correlation with DGESTAT
- The relationship between DGESTAT and DBIRWT appears to be nonlinear
- The propensity score is estimated using covariate information

Working MI Models

- Illustrative MI models focus on the effect of DBIRWT on predicting DGESTAT
- Model I: including all the covariates except DBIRWT
- Model II: model I plus including DBIRWT as a linear predictor
- Model III: model I plus including the quadratic term of DBIRWT
- Model IV: models I-III treat DGESTAT as continuous and the imputed values can be fractional numbers. Yet the original unit is integer (days). Implement the predictive mean matching version of model III so that the imputed values are all integers

Summery of Results

- Model I omits an important predictor, we see many flags (of imbalance) between the matched sample of observed and imputed data
- Model II significantly improves over model I, yet we still see some differences at the tail of DBIRWT. This is due to the fact that a linear predictor does not fully capture the nonlinear relationship between the two
- Model III and IV improve further, leave only one or two places flagged by the difference of the variance

Balancing Properties

- Under the strongly ignorable assumption

  $\Pr(Y_{mis} \,|X,R=0)=\Pr(Y_{obs} \,|X,R=1)$ (1)
- By the property of the propensity score

  $\Pr(X|g(X),R=0)=\Pr(X|g(X),R=1)$ (2)
- However, Eq. (1) has a limited practical use if X is multidimensional
- With Eq. (2), Eq. (1) is equivalent to

  $\Pr(Y_{mis} \,|g(X),R=0)=\Pr(Y_{obs} \,|g(X),R=1)$ (3)

  $\Pr(X|Y_{mis} \,, g(X),R=0)=\Pr(X|Y_{obs,} \, g(X),R=1)$ (4)
- From Eq. (1) to Eqs. (3) and (4), a p-dimensional problem is reduced to 1 and 2-dimensional problems

Conclusion

- The balancing relationships also hold if g(X) is replaced by a more general balancing score b(X)
- How to connect the diagnostic results to post-imputation inference
- Simulation studies demonstrate the utility of the proposed diagnostic strategy
- Extend to multivariate missing data situations in the future