Measurement, Design, and Analytic Techniques in Mental Health and Behavioral Sciences

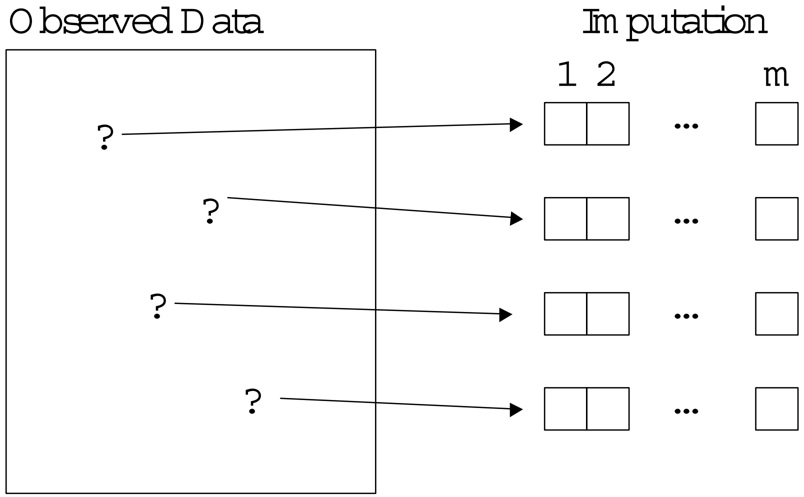
Lecture 7 (January 25, 2007): Multiple Imputation

XH Andrew Zhou

azhou@u.washington.edu

Professor, Department of Biostatistics, University of Washington

Multiple imputation



Multiple imputation, cont

- Imputation: Create D imputations of the missing data, $Y_{mis}^{(1)}, \ldots, Y_{mis}^{(D)}$, under a suitable model.
- Analysis: Analyze each of the D completed data sets in the same way.
- Combination: Combine the D sets of estimates and SE's using Rubin's (1987) rules.

Multiple imputation, cont

The advantages in using multiple imputation techniques:

- Allow the use of simple complete-data techniques and software
- The data collector (the imputer) and the data analyst may be different
- Reflect the sampling variability that occur due to the missing values
- Reflect uncertainty of the model if the imputations are drawn from different models
- One set of imputations may be used for many analyses.
- Highly efficient even for very small D.

The MI disadvantage

- requires more work

Efficiency

- The efficiency (on the variance scale) of an estimator of the scalar parameter based on D imputations to one based on an infinite number of imputations is approximately

$$(1 + \frac{\lambda}{D})^{-1}.$$

- Here λ is the fraction of missing information.

Efficiency (%)

	λ				
D	0.1	0.3	0.5	0.7	0.9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

How MI works

Three phases:

- Create imputations
- Analyze the imputed data sets
- Combine the results

Analysis step

- Analyze each imputed data set in the same way using complete-data methods.
- Store D sets of point estimates and standard errors

Combining the results

For a scalar parameter θ (Rubin, 1987)

- $(\widehat{\theta}_d, V_d)$: point estimates and variance estimates for dth imputed data set.
- MI point estimate: $\bar{\theta} = \frac{1}{D}\sum_{d=1}^{D}\widehat{\theta}_d$.
- Within variance: $\bar{V} = \frac{1}{D}\sum_{d=1}^{D} V_d$.
- Between variance: $B = \frac{1}{D-1}\sum_{d=1}^{D}(\widehat{\theta}_d - \bar{\theta})^2$
- Total variance: $T = \bar{V} + (1 + D^{-1})B$

Theoretical justification on multiple imputation

Large sample Bayesian Approximation

- Using iterative procedures, we create draws from the posterior distribution of θ
- In that case a large number of draws are needed
- If we assume normality of the observed-data posterior distribution, we need to estimate only the mean and variance–much less draws are needed
- In that case, a very limited number of draws are required to estimate reliably the distribution mean.
- The MI is based on this idea

Justification, cont

- If we assume $p(\theta \mid y_{obs})$ is approximately normal, the observed-data posterior can be effectively determined by the posterior mean and variance, $E(\theta \mid y_{obs})$ and $Var(\theta \mid Y_{obs})$.
- Note that

$$E(\theta \mid y_{obs}) = E[E(\theta \mid y_{mis}, y_{obs}) \mid Y_{obs}] =$$

$$\int E(\theta \mid y_{mis}, y_{obs} p(y_{mis} \mid y_{obs}) dy_{mis},$$

where the outer expectation is taken with respect to the posterior predictive distribution, $p(y_{mis} \mid y_{obs})$.

Justification, cont

- $$Var(\theta \mid Y_{obs}) = E[Var(\theta \mid Y_{mis}, Y_{obs}) \mid Y_{obs}] +$$
  $$Var[E(\theta \mid Y_{mis}, Y_{obs}) \mid Y_{obs}],$$
  where the outer expectation and variance are taken with respect to $p(y_{mis} \mid y_{obs})$.
- $$E[Var(\theta \mid Y_{mis}, Y_{obs}) \mid Y_{obs}] =$$
  $$\int Var(\theta \mid Y_{mis}, Y_{obs}) p(Y_{mis} \mid Y_{obs}) dY_{mis}.$$
- $$Var[E(\theta \mid Y_{mis}, Y_{obs}) \mid Y_{obs}] =$$
  $$\int E^2(\theta \mid Y_{mis}, Y_{obs}) p(Y_{mis} \mid Y_{obs}) dY_{mis} - (\int E(\theta \mid Y_{mis}, Y_{obs}) p(Y_{mis} \mid Y_{obs}))^2.$$

Justification, cont

For large D,

- $$E[E(\theta \mid y_{mis}, y_{obs})y_{obs}] \approx \frac{1}{D}\sum_{d=1}^{D} \hat{\theta}_d,$$

  where $y_{mis}^{(d)}$ are independent draws of y_{mis} from the posterior predictive distribution, $p(y_{mis} \mid y_{obs})$, and $\hat{\theta}_d = E(\theta \mid y_{mis}^{(d)}, y_{obs})$, the complete-data posterior mean of θ calculated for the dth imputed data set $(y_{mis}^{(d)}, y_{obs})$.

Justification, cont

- $$E[Var(\theta \mid y_{mis}, y_{obs}) \mid y_{obs}] \approx$$

  $$\bar{V} = \frac{1}{D} \sum_{d=1}^{D} Var(\theta \mid y_{mis}^{(d)}, y_{obs}),$$

  where $Var(\theta \mid y_{mis}^{(d)}, y_{obs})$ is the complete-data posterior variance of θ calculated for the dth imputed data set $(y_{mis}^{(d)}, y_{obs})$, and

- $$Var[E(\theta \mid y_{mis}, y_{obs}) \mid y_{obs}] \approx$$

  $$B = \frac{1}{D-1} \sum_{d=1}^{D} (\hat{\theta}_d - \bar{\theta})^2,$$

  where $\bar{\theta} = \frac{1}{D} \sum_{d=1}^{D} \hat{\theta}_d$.

Justification, cont

- $\bar{V}$: within-imputation variance
- B: between-imputation variance

Justification for combining rule

- MI point estimate for $E(\theta \mid y_{obs})$ (that is, for θ):

$$\bar{\theta} = \frac{1}{D} \sum_{d=1}^{D} \hat{\theta}_d.$$

- MI estimate for $Var(\theta \mid y_{obs})$ is

$$\bar{V} + B,$$

which is good when the between variance is small.

- However, a better estimate for $Var(\theta \mid y_{obs})$ is

$$T = \bar{V} + (1 + D^{-1})B.$$

MI inferences on scalar θ

- A further refinement for small D is to replace the normal distribution by a t distribution for the statistics, $(\theta - \bar{\theta})/\sqrt{T}$. That is,

$$T^{-1/2}(\theta - \bar{\theta}) \sim t_\nu,$$

  with the degrees of freedom $\nu = (D-1)(1 + r_D^{-1})^2$,where $r_D = \frac{(1+D^{-1})B}{\bar{V}}$, the relative increase in variance due to missing data.

- When the completed data sets are based on limited degrees of freedom, say v_{com}, an additional refinement replaces ν with:

$$\nu^* = (v^{-1} + \hat{\nu}_{obs}^{-1})^{-1},$$

  where

$$\hat{\nu}_{obs} = (1 - r_D)\frac{\nu_{com} + 1}{\nu_{com} + 3}\nu_{com}.$$

  See Barnard and Rubin (1999, Biometrika) for detail.

MI inferences on scalar θ, cont

- A $100(1-\alpha)\%$ confidence interval for θ is

$$\bar{\theta} \pm t_{\nu,1-\alpha/2}\sqrt{T},$$

a p-value for testing the null hypothesis that $\theta = \theta_0$ against a two-sided alternative is

$$2P(t_\nu \geq T^{-1/2} \mid \bar{\theta} - \theta_0 \mid)$$

Or equivalently,

$$P(F_{1,\nu} \geq T^{-1}(\bar{\theta} - \theta)^2).$$

Missing information rate

- The estimated fraction of missing information about θ is given by

$$\hat{\lambda} = \frac{r_D + 2/(\nu + 3)}{r_D + 1}.$$

MI Estimation when θ is not scalar

- When θ is not a scalar but a vector with k dimensions, finding an adequate reference distribution for the statistic

$$(\bar{\theta} - \theta)' V(\theta \mid Y_{obs})^{-1} (\bar{\theta} - \theta)/k$$

  is not a simple matter.
- The main problem is that for small D, the between-imputation covariance matrix B is a very noisy estimate of $V(\theta \mid Y_{obs})$, and does not even have full rank if $D \leq k$.

Estimation when θ is not scalar, Cont

- One way out of this difficulty is to make the simplifying assumption that the population between- and within-imputation covariance matrices are proportional to one another which is equivalent to assuming that the factions of missing information for all components of θ are equal.
- Under this assumption, a more stable estimate of total variance is
$$\tilde{V}(\theta \mid Y_{obs}) = (1 - r_D)\bar{V},$$
where $r_D = (1 + D^{-1})tr(B\bar{V}^{-1})/k$ is the average relative increase in variance due to missing data across the components of θ, and $tr(B\bar{V}^{-1})$ is the trace of $B\bar{V}^{-1}$, the sum of main diagonal elements of $B\bar{V}^{-1}$.

Hypothesis testing when θ is not scalar, cont

Combining point estimates and covariance matrices:

- Then, under the null hypothesis $H_0 : \theta = \theta_0$, the test statistics

$$W(\theta_0, \theta) = (\theta_0 - \bar{\theta})^T \bar{V}^{-1} (\theta_0 - \bar{\theta})/(1 + r_D)k$$

  has a F-distribution with the degrees of freedom k and ν_1.

- Hence, the p-value=$P(F_{k,\nu_1} > W(\theta_0, \theta))$.

Hypothesis testing when θ is not scalar, cont

- Here the degree of freedom

$$\nu_1 = 4 + (k(D-1)-4)[1+\frac{a}{r_D}]^2, a = 1 - \frac{2}{k(D-1)}$$

  if $k(D-1) > 4$. When $k(D-1) \leq 4$,

$$\nu_1 = (k+1)\nu/2 = (k+1)(D-1)(1+r_D^{-1})^2/2.$$

- Although the above reference distribution is derived under the strong assumption that the fractions of missing information for all components of θ are equal, Li and Raghunanthan and Rubin (1991, JASA) reported encouraging results even when this assumption is violated.

Hypothesis testing when θ is not scalar, cont

- Assume there are nuisance parameters ϕ, in addition to the parameter of interest θ.
- Our null and alternative hypotheses are that $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.
- Let $\widehat{\theta}$ and $\widehat{\phi}$ be estimates of θ and ϕ without H_0, and let $\widehat{\phi}_0$ be estimates of ϕ under H_0 when there are no missing data.
- Then, The P value for $\theta = \theta_0$ based on the likelihood-ratio test will be $Pvalue = Pr(\chi^2_k > LR)$ where $LR = LR[(\hat{\theta}, \hat{\phi}), (\theta_0, \hat{\phi}_0)]$, and χ^2_k is a χ^2 random variable with k degrees of freedom.

Hypothesis testing when θ is not scalar, cont

Likelihood ratio test

- For the dth imputed data set $(y_{mis}^{(d)}, y_{obs})$, let $(\widehat{\theta}^{(d)}, \widehat{\phi}^{(d)})$ be the estimates of θ and ϕ without assuming H_0 and $\widehat{\phi}_0$ is an estimate of ϕ under $H_0 : \theta = \theta_0$, and $LR^{(d)}$ be the corresponding likelihood ratio test statistics.
- Let $\bar{\theta} = \sum_{d=1}^{D} \widehat{\theta}_d / D$, $\bar{\phi} = \sum_{d=1}^{D} \bar{\phi}^{(d)}$, $\bar{\phi}_0 = \sum_{d=1}^{D} \bar{\phi}_0^{(d)}$, and $\bar{LR} = \sum_{d=1}^{D} LR^{(d)} / D$.
- Assume that the function LR can be evaluated for each of the D completed data sets at $\bar{\theta}, \bar{\phi}$, θ_0, and $\bar{\phi}_0$ to obtain D values of $LR[(\bar{\theta}, \bar{\phi}), (\theta_0, \bar{\phi}_0)]$ whose average across the D imputations is $\bar{LR}_0$.

Hypothesis testing when θ is not scalar, cont

- Then the test statistics,

$$W = \bar{LR}_0/[k + \frac{(D+1)(\bar{LR} - \bar{LR}_0)}{(D-1)}]$$

  is identical in large samples to $W(\theta_0, \bar{\theta})$ and can be used exactly as it if it were $W(\theta_0, \bar{\theta})$ (Meng and Rubin, 1992, Biometrika).
- Hence, the p-value=$P(F_{k,\nu_1} > W)$.

Hypothesis testing when θ is not scalar, cont

In some cases, the complete-data method of analysis may not produce estimates of the general function $LR(.,.,.,)$ but only the value of the likelihood ratio statistic. So if we do not have $L\bar{R}_0$ but only $LR_1, \ldots, LR_D$, there is a less accurate way to combine this value (Li et al, 1991).

- The repeated-imputation P value is given by

$$P(F_{k,b} > L\tilde{R}),$$

where

$$L\tilde{R} = \frac{\frac{L\bar{R}}{k} - (1 - D^{-1})\nu}{1 + (1 + D^{-1}\nu},$$

ν is the sample variance of $(\sqrt{LR_1}, \ldots, \sqrt{LR_D})$, and

$$b = k^{-3/D}(D-1)\{1 + [(1 + D^{-1})\nu]^{-1}\}^2.$$

Practice guidelines - asymptotic consideration

- In MI, the rules for combining complete-data inferences all assume that sample is large enough for usual asymptotic approximation to hold.
- For smaller samples, when the asymptotic methods break down, simulation-based summaries of the posterior distribution of θ may be preferable, keeping in mind Bayesian interpretation depends on a prior.

Practice guidelines - rates of missing information

- When the rate of missing information is low, MI estimates based on, say, $D = 5$ imputations may be nearly precise as average over hundreds of draws of θ.
- With high rates of missing information, however, a larger number of imputations may be necessary.

Practice guidelines - robustness

- Parametric Bayesian simulation methods depends on heavily on the correct form of the parametric complete-data model.
- MIs created under a false model may not have a disastrous effect on the final inference, provided the analyses of imputed data sets are done under more plausible assumptions.

Choosing an imputation model

- Because the imputation and analysis steps are distinct, is it possible to have valid MI inferences if the imputer's model and the analyst's model are different?
- The rules for combining complete-data inferences were derived under some implicit assumptions of agreement between these two models.

More restrictive analyst’s model

- The analyst’s model is a special case of imputer’s one.
- If the analyst’s extra assumption is true, MI inferences will be valid, but may be conservative because the imputations will reflect an extra degree of uncertainty.
- If the analyst’s extra assumption is not true, MI inferences will be not valid.

More restrictive imputer's model

- The analyst's model is more general than the imputer's; that is, the imputer makes assumptions to the complete data that the analyst does not.
- If the imputer's extra assumption is true, MI inferences will be still valid.
- In addition, the MI estimate $\bar{\theta}$ is more efficient than an observed data estimate derived purely from the analyst's model, because the MI estimate incorporates the imputer;s superior knowledge about the data, a property called superefficiency.
- Moreover, the MI interval has average width that is shorter than a confidence interval derived based on the observed data and the analyst's model.
- If the analyst's extra assumption is not true, MI inferences will be not valid.

Imputation model

The imputation model should include

- variables crucial to the analysis
- variables that are highly predictive of the variables that are crucial to the analysis (e.g an outcome)
- variables that are highly predictive of missingness
- variables that describe special features of the sample design (probability surveys)

A general guideline is that the imputed should use a model that is general enough to preserve any associations among variables (two-, three-, or even higher-way associations) that may be the target of subsequent analyses.