Multiple Imputation, I

John M. Abowd

March, 2005

1 Models, Parameters, Bayes Law

1.0.1 Symmetric treatment of data and parameters (joint distribution of the data and the parameters)

$$p(Y, \theta)$$

1.0.2 Two factorizations

Likelihood and prior parameters:

$$p(Y, \theta) = p(Y|\theta)\, p(\theta)$$

where the first factor, $p(Y|\theta) \equiv \ell(\theta|Y)$ is the data model (or likelihood function) and the second factor $p(\theta)$ is the prior distribution of the parameters.

Alternatively, posterior parameters and marginal distribution of the data:

$$p(Y, \theta) = p(\theta|Y)\, p(Y)$$

where $p(\theta|Y)$ is the posterior distribution of the parameters (note that $p(\theta|Y)$ is not the same function as $\ell(\theta|Y)$) and $p(Y)$ is the marginal (predictive) distribution of the data. By direct maniupulation, we can derive the posterior distribution of the parameters

$$p(\theta|Y) = \frac{p(Y, \theta)}{p(Y)} = \frac{p(Y|\theta)\, p(\theta)}{p(Y)} \equiv \frac{\ell(\theta|Y)\, p(\theta)}{p(Y)}.$$

2 Bayes Rule with Missing Data

Now, consider what happens if we observe some of the data but not other parts. Partition Y as

$$Y = (Y_{mis}, Y_{obs})$$

and let the matrix M be defined by

$$m_{ij} = \begin{cases} 1, & \text{if } y_{ij} \text{ is missing} \\ 0, & \text{otherwise} \end{cases}$$

Then, the joint distribution of interest is

$$p(Y_{mis}, Y_{obs}, M, \theta, \psi) = p(Y_{mis}, Y_{obs}, M \,|\theta, \psi)\, p(\theta, \psi)$$

Decomposing leads to

$$p(Y_{mis}, Y_{obs}, M \,|\theta, \psi) = p(M \,|Y_{mis}, Y_{obs}, \psi)\, p(Y_{mis}, Y_{obs} \,|\theta)$$

where the parameters ψ are associated with the missing data mechanism and the parmeters θ are associated with the complete data model (or likelihood function). There can be a functional dependence between ψ and θ, for example $(\theta, \psi) = (\theta(\beta), \psi(\beta))$; so, there is no loss of generality in this notation. The distribution $p(M \,|Y_{mis}, Y_{obs}, \psi)$ is called the missing data generating mechanism. The distribution $p(Y_{mis}, Y_{obs} \,|\theta)$ is called the complete data model.

3 Ignorability (Likelihood based)

To do inference in the presence of Y_{mis} consider each necessary piece

$$\begin{aligned} p\left(Y_{obs}, M \left|\theta, \psi\right.\right) &= \int p\left(M \left|Y_{mis}, Y_{obs}, \psi\right.\right) p\left(Y_{mis}, Y_{obs} \left|\theta\right.\right) dY_{mis} \\ &= p\left(M \left|Y_{obs}, \psi\right.\right) \int p\left(Y_{mis}, Y_{obs} \left|\theta\right.\right) dY_{mis} \end{aligned}$$

Likelihood-based ignorability conditions hold:

- Data missing at random (MAR) $p\left(M \left|Y_{mis}, Y_{obs}, \psi\right.\right) = p\left(M \left|Y_{obs}, \psi\right.\right)$.
- Missing data and model parameters are distinct $(\theta, \psi) \neq \left(\theta\left(\beta\right), \psi\left(\beta\right)\right)$ for some lower dimensional space β.

Bayesian ignorability conditions:

- Data missing at random (MAR) $p\left(M \left|Y_{mis}, Y_{obs}, \psi\right.\right) = p\left(M \left|Y_{obs}, \psi\right.\right)$.
- The prior distribution of (θ, ψ) factors into $p\left(\theta\right) p\left(\psi\right)$.

4 Imputing Missing Data

Multiple imputation is based on the posterior predictive distribution of Y_{mis}.

$$p\left(Y_{mis}\left|Y_{obs},M,\theta,\psi\right.\right)=\frac{p\left(M\left|Y_{mis},Y_{obs},\psi\right.\right)p\left(Y_{mis},Y_{obs}\left|\theta\right.\right)}{p\left(Y_{obs},M\left|\theta,\psi\right.\right)}$$

Integrating out the parameters yields

$$p\left(Y_{mis}\left|Y_{obs},M\right.\right)=\iint\frac{p\left(M\left|Y_{mis},Y_{obs},\psi\right.\right)p\left(Y_{mis},Y_{obs}\left|\theta\right.\right)}{p\left(Y_{obs},M\left|\theta,\psi\right.\right)}p\left(\theta,\psi\left|Y_{obs},M\right.\right)d\theta d\psi$$

5 Operationalizing the Distribution

This distribution is the basis for multiple imputation of missing data. To be useful it must be operationalized considering the same kinds of assumptions that are discussed above. Imposing MAR and Bayesian ignorability yields:

$$
\begin{aligned}
p\left(Y_{mis}\left|Y_{obs}, M\right.\right) &= E_{(\theta,\psi|Y_{obs},M)}\left[\frac{\text{Joint Distribution}\left(Y_{mis}, Y_{obs}, M\right)}{\text{Joint Distribution}\left(Y_{obs}, M\right)}\right] \\
&= \iint\left[\frac{p\left(M\left|Y_{mis}, Y_{obs}, \psi\right.\right) p\left(Y_{mis}, Y_{obs}\left|\theta\right.\right)}{p\left(Y_{obs}, M\left|\theta, \psi\right.\right)}\right] p\left(\theta, \psi\left|Y_{obs}, M\right.\right) d\theta d\psi \\
&= \iint\frac{p\left(M\left|Y_{obs}, \psi\right.\right) p\left(Y_{mis}, Y_{obs}\left|\theta\right.\right)}{p\left(M\left|Y_{obs}, \psi\right.\right)\int p\left(Y_{mis}, Y_{obs}\left|\theta\right.\right) dY_{mis}} p\left(\theta\left|Y_{obs}\right.\right) p\left(\psi\left|Y_{obs}, M\right.\right) d\theta d\psi \\
&= \int\frac{p\left(Y_{mis}, Y_{obs}\left|\theta\right.\right)}{\int p\left(Y_{mis}, Y_{obs}\left|\theta\right.\right) dY_{mis}} p\left(\theta\left|Y_{obs}\right.\right) d\theta \int p\left(\psi\left|Y_{obs}, M\right.\right) d\psi \\
&= p\left(Y_{mis}\left|Y_{obs}\right.\right)
\end{aligned}
$$

6 Details

$$p\left(Y_{obs}, M\,|\theta, \psi\right) = p\left(M\,|Y_{obs}, \psi\right) \int p\left(Y_{mis}, Y_{obs}\,|\theta\right) dY_{mis}$$

$$p\left(Y_{mis}\,|Y_{obs}\;\theta\right) = \frac{p\left(Y_{mis}, Y_{obs}\,|\theta\right)}{\int p\left(Y_{mis}, Y_{obs}\,|\theta\right) dY_{mis}}$$

and

$$\int p\left(\psi\,|Y_{obs}, M\right) d\psi = 1$$

Note, there are other ways to operationalize the posterior predictive distribution so that ignorability does not have to be imposed.

7 The Bayesian Bootstrap Mulitple Imputation Procedure

Let (Y_{obs}, Y_{mis}) be partitioned such that observations $1, \ldots, n_{obs}$ correspond to the data in Y_{obs} and $n_{obs}+1, \ldots, n$ correspond to the data in Y_{mis}.

Algorithm:

- Draw $n_{obs} - 1$ random numbers from $U(0,1)$. Sorted from lowest to highest labeled $a_1, \ldots a_{n_{obs}-1}$. Define $a_0 = 1$ and $a_{n_{obs}} = 1$.
- Impute each of $n - n_{obs} = n_{mis}$ values of Y_{mis} by sampling with replacement from Y_{obs} using the probabilities $(a_1 - a_0), \ldots, (a_{n_{obs}} - a_{n_{obs}-1})$. I.e., draw n_{mis} random numbers from $U(0,1)$; impute Y_i when $a_{i-1} < u \leq a_i$.

The Bayesian Bootstrap (BB) preserves all multivariate relations among the imputed variables.

The BB can be modified to accomodate non-monotone missing data.

8 The Sequential Regression Multivariate Imputation Procedure

Impose Bayesian ignorability

Estimate $p\left(Y_{mis}|Y_{obs}\right) = \int p\left(Y_{mis}|Y_{obs},\theta\right) p\left(\theta|Y_{obs}\right) d\theta$.

Sample M times from $p\left(Y_{mis}|Y_{obs}\right)$ as follows

$$p\left(y_1, y_2, ..., y_K|\theta_1, \theta_2, ..., \theta_K\right) = p_1\left(y_1|\theta_1\right) p_2\left(y_2|y_1, \theta_2\right) ... p_K\left(y_K|y_1, y_2, ..., y_{K-1}, \theta_K\right)$$

This equation always holds and is approximated by the SRMI procedure.

The SRMI imputation procedure consists of L rounds. Denote the completed data in round $\ell + 1$ on some variable y_k by $y_k^{(\ell+1)}$. In round $\ell + 1$, missing values of y_k are drawn from the predictive density corresponding to the conditional density:

$$p_k\left(y_k|y_1^{(\ell+1)}, y_2^{(\ell+1)}, ..., y_{k-1}^{(\ell+1)}, y_{k+1}^{(\ell)}, ..., y_k^{(\ell)}, \theta_k\right) \tag{1}$$

where the conditional density p_k is specified by an appropriate generalized linear model, and θ_k are the parameters of that model. Hence under SRMI, at each round ℓ, the variable under imputation is regressed on all non-missing data and the most recently imputed values of missing data. The imputation procedure stops after a predetermined number of rounds or when the imputed values are stable. Repeating the procedure M times yields M multiply-imputed data sets.

9 General Bayesian Methods for Multiple Imputation

9.0.3 Data Augmentation

These methods use the ignorable missing data hypothesis

$$p\left(\theta \left| Y_{obs}, M \right.\right) = cons \times p\left(\theta\right) \times \ell\left(\theta \left| Y_{obs} \right.\right)$$

which means that we do not have to estimate $p\left(M \left| Y_{obs}, \psi \right.\right)$.

Algorithm for $\ell = 0, \ldots$. Let $Y_{mis}^{(0)} = 0$ and $\theta^{(0)} = 0$

- I-step: Draw $Y_{mis}^{(\ell+1)}$ from $p\left(Y_{mis} \left| Y_{obs}, \theta^{(\ell)} \right.\right)$
- P-step: Draw $\theta^{(\ell+1)}$ from $p\left(\theta \left| Y_{obs}, Y_{mis}^{(\ell)} \right.\right)$

As $\ell \rightarrow \infty$, $p^{(\ell)}\left(Y_{mis}, \theta \left| Y_{obs} \right.\right) \rightarrow p\left(Y_{mis}, \theta \left| Y_{obs} \right.\right)$, the joint posterior distribution of the missing data and the parameters.

10 Assessing the Inference Validity of Imputation Procedures

It is helpful now to specify some variables as complete, X, and other variables as potentially missing, Y. The population consists of N entities with data (X, Y), where X is $N \times q$; Y is $N \times p$.

A variable $Y_{.j}$ is included in the sample (design) as indicated by the column $I_{.j}$. So I is $N \times p$ and indicates all of the sampled entites across all variables. (Recall that complete X means that we have a well defined frame population will all frame data complete.)

A variable $Y_{.j}$ is observed as indicted by the column $R_{.j}$. So R is $N \times p$ and indicates all of the responses to all the surveyed variables.

The term survey can be replaced with record when using an administrative data base for which a frame population is available.

11 Sampling and Missing Data

X is always observed in all samples.

$Y = (Y_{inc}, Y_{exc})$ and $R = (R_{inc}, R_{exc})$ according to I. An entity i has been sampled for variable j according to $I_{ij} = 1$.

Y_{inc} is further decomposed into (Y_{obs}, Y_{mis}) according to R. An entity i is observed for variable j according to $R_{ij} = 1$.

$Y = (Y_{obs}, Y_{mis}, Y_{exc}) = (Y_{obs}, Y_{nob})$, where $Y_{nob} = (Y_{mis}, Y_{exc})$

12 Estimands and Their Posterior Distributions

An estimand is some function of the population data that we wish to estimate, say $Q(X, Y)$.

$$P(Q \,|X, Y_{obs}, R_{inc}, I) = \int_{\{Y_{nob}|Q(X,Y)=Q'\}} P(Y_{nob} \,|X, Y_{obs}, R_{inc}, I)\, dY_{nob}$$

and

$$P(Y_{nob} \,|X, Y_{obs}, R_{inc}, I) = \frac{\int P(X, Y)\, P(R \,|X, Y)\, P(I \,|X, Y, R)\, dR_{exc}}{\iint P(X, Y)\, P(R \,|X, Y)\, P(I \,|X, Y, R)\, dR_{exc} dY_{nob}}$$

which can, in general, be evaluated from the properties of the sample design and the assumptions made about the missing data mechanism, e.g., MAR.

13 Inference Validity with Missing Data

13.1 Population Quantities

Population value of the estimand and a measure of its variability:

$$Q_0(X, Y) \text{ and } U_0(X, Y)$$

In the complete finite population, these values are given.

13.1.1 Complete Data Statistics

- Estimand: $\hat{Q}(X, Y_{inc}, I)$
- Precision/Variance measure: $\hat{U}(X, Y_{inc}, I)$

13.1.2 Multiple Imputation Statistics

- Estimand: $\bar{Q}_L(X, Y_{obs}, I, R_{inc})$, based on L multiple imputations
- Variance measure (within): $\bar{U}_L(X, Y_{obs}, I, R_{inc})$, based on L multiple imputations
- Variance measure (between): $B_L(X, Y_{obs}, I, R_{inc}) = \frac{1}{L-1} \sum_{\ell} \left(\bar{Q}_\ell - \bar{Q}_L\right) \left(\bar{Q}_\ell - \bar{Q}_L\right)'$
- Total Variance: $T_L = \bar{U}_L(X, Y_{obs}, I, R_{inc}) + \left(1 + \frac{1}{L}\right) B_L(X, Y_{obs}, I, R_{inc})$

14 Complete Data Must Be Randomization Valid

If we sample with probability distribution $P\left(I\left|X,Y\right.\right)=P\left(I\left|X\right.\right)$,i.e., the sample design depends only on the data X, for which we have a complete frame population, then the inference system is randomization valid if

$$\left(\hat{Q}\left|X,Y\right.\right)\sim N(Q_0\left(X,Y\right),U_0\left(X,Y\right))$$

and

$$\left(\hat{U}\left|X,Y\right.\right)\sim\left(U_0\left(X,Y\right),\ll U_0\left(X,Y\right)\right)$$

where $\ll$ means that the matrix tends to 0 as the finite population approaches ∞.

We assume that the estimand has these properties in complete data so that we can focus on when the imputation system is randomization valid.

15 Randomization Validity of Infinite-L Repeated Multiple Imputation

A multiple imputation procedure is randomization valid for infinite L if

$$\left(\bar{Q}_\infty \left| X, Y \right.\right) \sim N(Q_0 \left(X, Y\right), T_0 \left(X, Y\right))$$

and

$$\left(T_\infty \left| X, Y \right.\right) \sim \left(T_0 \left(X, Y\right), \ll T_0 \left(X, Y\right)\right)$$

where $\bar{Q}_\infty$ is $\bar{Q}_L(X, Y_{obs}, I, R_{inc})$ as $L \longrightarrow \infty$, and similarly for $T_\infty = U_\infty + B_\infty$.

When the conditions for randomization validity hold, the multiple imputation procedure provides a sampling distribution for $\hat{Q}_L$ that is asymptotically correct as $L \longrightarrow \infty$:

$$\left(\hat{Q}_L - \bar{Q}_\infty \left| X, Y_{obs}, I, R_{inc} \right.\right) \sim N\left(0, T_\infty\right)$$

which allows one to make probability statements about hypothesized values of Q in the usual manner.