A Note on Testing for the Type of Missingness

Jude C. Hays
Michaël Aklin
Junghyun Lim

University of Pittsburgh

March 22, 2019
Political methodologists have revisited recently the wisdom of multiple imputation as an alternative to listwise deletion. Multiple imputation weakly dominates listwise deletion when missingness, conditional on the observed data, is random (MAR). However, when missingness is nonignorable (NI) the decision of whether to multiply impute missing data (MI) or listwise delete cases with missing data (LD) may be more complicated. Focusing on outcomes, we propose a simple balance test to help distinguish random missingness from non-ignorable missingness. To implement this test it is necessary separate variables \textit{ex ante} into three categories: variables that provide information about the outcome only, variables that provide information about missingness only (i.e., proxy variables), and variables that provide information about both missingness and outcomes. Theory should guide the separation of variables into the different categories, but these beliefs can be bolstered with empirical evidence. Our Monte Carlo results show the balance test has good small sample properties and that MI with high quality proxies is preferable to LD.
1. INTRODUCTION

Political methodologists have revisited recently the wisdom of multiple imputation as an alternative to listwise deletion (Arel-Bundock and Pelc, 2018; Pepinsky, 2018). Multiple imputation weakly dominates listwise deletion when missingness, conditional on the observed data, is random (MAR). However, when missingness is nonignorable (NI) the decision of whether to multiply impute missing data (MI) or listwise delete cases with missing data (LD) may be more complicated. Focusing on outcomes, we propose a simple balance test to help distinguish random missingness from non-ignorable missingness. To implement this test it is necessary separate variables \textit{ex ante} into three categories: variables that provide information about the outcome only, variables that provide information about missingness only (i.e., proxy variables), and variables that provide information about both missingness and the outcome. Theory should guide the separation of variables into the different categories, but these beliefs can be bolstered with empirical evidence. Our Monte Carlo results show the balance test has good small sample properties and that MI with high quality proxies is preferable to LD.

2. A BALANCE TEST FOR NONIGNORABLE MISSINGNESS

Consider the following DGP:

\[ y = \phi_1 x + \phi_2 z + \epsilon \]
\[ x = \phi_{11} X_1 + \phi_{12} X_2 \]
\[ z = \phi_{21} X_1 + \phi_{22} X_3 \]
\[ X_1, X_2, X_3, \epsilon \sim N(0, \sigma^2) \]  
(1)

\[ p = y + \eta \]
\[ \eta \sim N(0, \sigma^2_\eta) \]  
(2)

In this system, \( x \) is a \( n \times 1 \) vector of realized values from \( X \), a compound random variable that provides information about the outcome only; \( z \) is a \( n \times 1 \) vector of random variates from \( Z \), a compound random a variable that provides information about both missingness and the outcome. Among the components of \( X \) and \( Z \), \( X_1 \) is common to both \( X \) and \( Z \) while \( X_2 \) and \( X_3 \) are unique to \( X \) and \( Z \) respectively. All three components are independent and identically distributed. \( P \) is a random variable that provides information about missingness only. More specifically, \( P \) is a proxy variable that provides a noisy measure of the outcome \( Y \) (Achen, 1985). The variance of \( \eta \) determines the proxy’s signal-to-noise ratio.

The intuition for the test is as follows. If missingness on \( Y \) is random (MAR)—in this case, associated with \( Z \)—the linear expectation of \( X \) conditional on \( Z \) will differ across the full sample \((f)\) and post-listwise deletion subsample \((s)\) by chance only. Under MAR, the two samples will be balanced with respect to \( E(X|Z) \). If we reject the hypothesis that differences in the linear expectation of \( X \) conditional on \( Z \) are due to chance, the alternative is that missingness on \( Y \) is nonignorable (NI). NI missingness of \( Y \) can arise for two different reasons. First, missingness can be due to unobservables (NI type-I). Second, missingness can be due to the outcome itself (NI type-II). Consider the case of a single \( X \) and \( k - 1 \) variables that meet our definition for \( Z \):

\[ E(X|z_1, f \ldots z_{k-1}, f) = \beta_{0,f} + \beta_{1,f} z_1,f + \ldots + \beta_{k-1,f} z_{k-1},f \]
\[ E(X|z_1, s \ldots z_{k-1}, s, s = 1) = \beta_{0,s} + \beta_{1,s} z_1,s + \ldots + \beta_{k-1,s} z_{k-1},s . \]  
(3)

Allison (2001) demonstrates the logic of the test more formally. We can rewrite the two conditional expectation functions (for a single \( X \) and \( Z \)) as

\[ E(X|Z = z) = \int xf(x|z)dx \]
\[ E(X|Z = z, s = 1) = \int xf(x|z, s = 1)dx \]
These two functions are equivalent if \( f(x|z) = f(x|z, s = 1) \), and this equality will hold as long as the random variable \( X \) provides no information about missingness, or, more formally, \( \Pr(s = 1|x, z) = \Pr(s = 1|z) \).

\[
f(x|Z = z, s = 1) = \frac{f(x, z|s = 1)}{f(z|s = 1)} = \frac{\Pr(s = 1|x, z)f(x|Z = z)f(z)}{\Pr(s = 1|z)f(z)} = f(x|Z = z)
\]

The test evaluates the equality of coefficients across the two samples. This can be done using a Wald statistic,

\[
W = R(\beta)' \left[ \frac{\partial R(\beta)' \Sigma \partial R(\beta)}{\partial \beta} \right]^{-1} R(\beta),
\]

in which \( R(\beta) \) is a \( k \times 1 \) vector of constraints

\[
R(\beta) = \begin{bmatrix}
\beta_{0,f} - \beta_{0,s} \\
\beta_{1,f} - \beta_{1,s} \\
\vdots \\
\beta_{k,f} - \beta_{k,s}
\end{bmatrix},
\]

\( \frac{\partial R(\beta)}{\partial \beta} \) is the \( 2k \times k \) Jacobian matrix

\[
\frac{\partial R(\beta)}{\partial \beta} = \begin{bmatrix}
\frac{\partial (\beta_{0,f} - \beta_{0,s})}{\partial \beta_{0,f}} & \frac{\partial (\beta_{0,f} - \beta_{0,s})}{\partial \beta_{0,s}} & \cdots & \frac{\partial (\beta_{0,f} - \beta_{k,s})}{\partial \beta_{0,f}} & \frac{\partial (\beta_{0,f} - \beta_{k,s})}{\partial \beta_{0,s}} \\
\frac{\partial (\beta_{0,f} - \beta_{0,s})}{\partial \beta_{1,f}} & \frac{\partial (\beta_{0,f} - \beta_{0,s})}{\partial \beta_{1,s}} & \cdots & \frac{\partial (\beta_{0,f} - \beta_{k,s})}{\partial \beta_{1,f}} & \frac{\partial (\beta_{0,f} - \beta_{k,s})}{\partial \beta_{1,s}} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\frac{\partial (\beta_{0,f} - \beta_{0,s})}{\partial \beta_{k,f}} & \frac{\partial (\beta_{0,f} - \beta_{0,s})}{\partial \beta_{k,s}} & \cdots & \frac{\partial (\beta_{0,f} - \beta_{k,s})}{\partial \beta_{k,f}} & \frac{\partial (\beta_{0,f} - \beta_{k,s})}{\partial \beta_{k,s}}
\end{bmatrix},
\]

and \( \Sigma \) is the \( 2k \times 2k \) variance-covariance matrix of coefficients

\[
\Sigma = \begin{bmatrix}
\text{var} (\beta_{0,f}) & \text{cov} (\beta_{0,f}, \beta_{1,f}) & \cdots & \text{cov} (\beta_{0,f}, \beta_{k,f}) & \text{cov} (\beta_{0,f}, \beta_{0,s}) & \text{cov} (\beta_{0,f}, \beta_{1,s}) & \cdots & \text{cov} (\beta_{0,f}, \beta_{k,s}) \\
\text{cov} (\beta_{1,f}, \beta_{0,f}) & \text{var} (\beta_{1,f}) & \cdots & \text{cov} (\beta_{1,f}, \beta_{k,f}) & \text{cov} (\beta_{1,f}, \beta_{0,s}) & \text{cov} (\beta_{1,f}, \beta_{1,s}) & \cdots & \text{cov} (\beta_{1,f}, \beta_{k,s}) \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\text{cov} (\beta_{k,f}, \beta_{0,f}) & \text{cov} (\beta_{k,f}, \beta_{1,f}) & \cdots & \text{var} (\beta_{k,f}) & \text{cov} (\beta_{k,f}, \beta_{0,s}) & \text{cov} (\beta_{k,f}, \beta_{1,s}) & \cdots & \text{cov} (\beta_{k,f}, \beta_{k,s}) \\
\text{cov} (\beta_{0,s}, \beta_{0,f}) & \text{cov} (\beta_{0,s}, \beta_{1,f}) & \cdots & \text{cov} (\beta_{0,s}, \beta_{k,f}) & \text{var} (\beta_{0,s}) & \text{cov} (\beta_{0,s}, \beta_{1,s}) & \cdots & \text{cov} (\beta_{0,s}, \beta_{k,s}) \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\text{cov} (\beta_{k,s}, \beta_{0,f}) & \text{cov} (\beta_{k,s}, \beta_{1,f}) & \cdots & \text{cov} (\beta_{k,s}, \beta_{k,f}) & \text{cov} (\beta_{k,s}, \beta_{0,s}) & \text{var} (\beta_{k,s}) & \cdots & \text{cov} (\beta_{k,s}, \beta_{1,s}) \\
\end{bmatrix}.
\]

The asymptotic variance-covariance matrix \( \Sigma \) is estimated using

\[
\Sigma = D(\beta)^{-1} \left( \sum_{j} w_{j} u_{j} u_{j}' \right) D(\beta)^{-1},
\]

in which, \( D(\beta) \) is the \( 2k \times 2k \) Jacobian of the gradient of the disturbance function, with two \( k \times k \) blocks along the diagonal, one for the full sample (f) and one for the post-listwise deletion subsample (s).
The gradient is a $1 \times 2k$ vector
\[
G(\beta) = \begin{bmatrix} G_f(\beta_0) & G_f(\beta_1) & \cdots & G_f(\beta_k) & G_s(\beta_0) & G_s(\beta_1) & \cdots & G_s(\beta_k) \end{bmatrix}.
\]

$u_j$ are the scores of $G(\beta)$ for observation $j$; and $w_j$ is a weight that takes a value of one for all observations that belong to both the full sample and post-listwise deletion subsample and takes a value of zero otherwise.

Under the null hypothesis $H_0: R(\beta) = 0$ (MAR), $W \xrightarrow{d} \chi^2(k)$. Rejecting the null hypothesis in favor of the alternative hypothesis $H_a: R(\beta) \neq 0$ implies rejecting MAR in favor of NI missingness.

Note that the Wald test based on the conditional expectation functions in (3) can be extended straightforwardly in two ways. First, the test can accommodate multiple $X$, variables that provide information about the outcome only. This creates a set of conditional expectation functions that are compared across the full sample and the listwise-delete subsample. Second, the test can be supplemented by including the conditional expectation function for the disturbance, $E(\varepsilon|z_1,s, \ldots z_l,s)$ and $E(\varepsilon|z_1,s, \ldots z_l,s, s = 1)$. Under the null hypothesis MAR, the parameters of these conditional expectation functions should differ by chance only across the full sample and list-wise deletion subsample.

We can also test for balance in the unconditional expectation functions, $E(X)$ and $E(X|s = 1)$. This amounts to a test of $\beta_{0,f} = \beta_{0,s}$. The unconditional expectations balance test provides empirical evidence to help separate $X’s$ and $Z’s$. $E(X)$ will be balanced across the full sample and listwise deletion subsample, while $E(Z)$ will not be balanced across these samples. Also, given that all covariates will be balanced if missingness is completely at random (MCAR), the unconditional balance test can be used to evaluate this possibility as well.

It is possible that $\Sigma$ is estimated poorly in small samples, and this would affect the performance of our balance test. Therefore, in the next section, we evaluate the small samples properties of the test. Our results show good performance with respect to both size and power. Moreover, in our experiments, MI with informative proxies is always preferable to LD, regardless of whether missingness is MAR or NI.

3. MONTE CARLO EXPERIMENTS

We use the DGP in equations (1) and (2) for our Monte Carlo experiments. To begin, all of the parameters—$\phi_1$, $\phi_2$, $\phi_{11}$, $\phi_{12}$, $\phi_{21}$, $\phi_{22}$, $\sigma^2$ and $\sigma_\varepsilon^2$—are set to one. The sample size for the full sample is 200, $n = 200$. Later, we manipulate $\sigma_\varepsilon^2$ to assess the relative performance of MI and LD under varying levels of informativeness for the proxy variable, $P$. We consider two levels of missingness (25% and 42% missingness) under MAR, NI type-I and NI type-II conditions. This gives six basic experiments: 25% MAR, 42% MAR, 25% NI type-I, 42% NI type-I, 25% NI type-II, and 42% NI type-II. With selection on Z and $\varepsilon$ (NI type-I), 25% missing implies $Y$ is missing when both $Z$ and $\varepsilon$ are below their respective 50th percentiles. 42% missing implies $Y$ is missing when both $Z$ and $\varepsilon$ are below their respective 65th percentiles.

Figure 1 shows the performance of our conditional expectations balance test under 42% MAR. (The results for 25% MAR are nearly identical.) We test for balance in $X$ conditioning on $Z$ only and then both $z$ and $\varepsilon$. The figure shows that the p-values provided from a $\chi^2$ distribution are accurately sized. In other words, the empirical distributions of the test statistic across repeated samples ($n = 200$) follow a $\chi^2$ distribution.
Figure 1: Testing for Balance in Conditional Expectations (42% MAR)

Note: With selection on Z, 25% missing implies all values of Y for which Z is below its 25th percentile are missing. 42% missing implies all values of Y for which Z is below its 42nd percentile are missing.

Table 1 compares the relative performance of MI and LD for the parameters $\phi_1$ and $\phi_2$ under MAR. This table confirms what we know about MAR. MI has bias and efficiency advantages over LD, although in this particularly simulation the advantages are slight.

Table 1: Bias, Overconfidence and RMSE Results for Selection on Z (MAR)

<table>
<thead>
<tr>
<th></th>
<th>bias($\phi_1$)</th>
<th>bias($\phi_2$)</th>
<th>s.e.($\phi_1$)</th>
<th>s.e.($\phi_2$)</th>
<th>r.m.s.e.($\phi_1$)</th>
<th>r.m.s.e.($\phi_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25% Missing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listwise Deletion</td>
<td>.001</td>
<td>-.001</td>
<td>1.01</td>
<td>.961</td>
<td>.068</td>
<td>.083</td>
</tr>
<tr>
<td>Multiple Imputation</td>
<td>-.001</td>
<td>-.001</td>
<td>.987</td>
<td>.971</td>
<td>.063</td>
<td>.074</td>
</tr>
<tr>
<td>42% Missing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listwise Deletion</td>
<td>.000</td>
<td>-.007</td>
<td>1.00</td>
<td>1.01</td>
<td>.077</td>
<td>.112</td>
</tr>
<tr>
<td>Multiple Imputation</td>
<td>.000</td>
<td>-.005</td>
<td>1.01</td>
<td>1.05</td>
<td>.071</td>
<td>.097</td>
</tr>
</tbody>
</table>

Note: With selection on Z, 42% missing implies all values of Y for which Z is below its 42nd percentile are missing.

Figure 2 shows the performance of our conditional expectations balance test under 42% NI type-I. (Again, the results for 25% MAR are nearly identical.) We test for balance in X conditioning on Z only and then both z and $\varepsilon$. In this case, selection on unobservables, we need the conditional expectation function for $\varepsilon$ to detect the nonignorable nature of the process driving missingness. The figure shows that under the alternative hypothesis (NI), the joint test has very good power, correctly rejecting the null hypothesis when it is false nearly 100% of the time.
Figure 2: Testing for Balance in Conditional Expectations (42% NI)

Note: With selection on Z and ε, 42% missing implies all values of Y for which both Z and ε are below their respective 65th percentiles are missing.

Table 2 compares the relative performance of MI and LD for the parameters $\phi_1$ and $\phi_2$. Given our DGP, $\sigma_2^2$ implies a very informative proxy with a linear correlation coefficient $\rho_{YY} = .91$. Given the strength of the signal, it is not too surprising that MI outperforms LD on every dimension: bias, standard error accuracy and the expected size of estimation errors.

Table 2: Bias, Overconfidence and RMSE Results for Selection on Z and ε (NI type-I)

<table>
<thead>
<tr>
<th></th>
<th>$\phi_1$</th>
<th>$\phi_2$</th>
<th>s.e.($\phi_1$)</th>
<th>s.e.($\phi_2$)</th>
<th>r.m.s.e.($\phi_1$)</th>
<th>r.m.s.e.($\phi_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25% Missing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listwise Deletion</td>
<td>-.001</td>
<td>-.217</td>
<td>1.02</td>
<td>.893</td>
<td>.064</td>
<td>.224</td>
</tr>
<tr>
<td>Multiple Imputation</td>
<td>-.001</td>
<td>-.116</td>
<td>.998</td>
<td>.932</td>
<td>.06</td>
<td>.13</td>
</tr>
<tr>
<td>42% Missing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listwise Deletion</td>
<td>.003</td>
<td>-.298</td>
<td>1.05</td>
<td>.957</td>
<td>.073</td>
<td>.306</td>
</tr>
<tr>
<td>Multiple Imputation</td>
<td>.002</td>
<td>-.161</td>
<td>1.04</td>
<td>.976</td>
<td>.068</td>
<td>.174</td>
</tr>
</tbody>
</table>

Note: With selection on Z and ε, 25% missing implies Y is missing when both Z and ε are below their respective 50th percentiles. 42% missing implies Y is missing when both Z and ε are below their respective 65th percentiles.

Figure 4 shows the performance of our conditional expectations balance test under 42% NI type-II. Again, we test for balance in X conditioning on Z only and then both z and ε. With selection on the outcome, both
tests have good power against the Null hypothesis when it is false.

Figure 3: Testing for Balance in Conditional Expectations (42% NI II)

![Graph showing empirical distribution of test statistic against theoretical PDF for chi-squared distribution](image)

**Note:** With selection on $Y$, 42% missing implies all values below its 42nd percentile are missing.

Table 3 compares the relative performance of MI and LD for the parameters $\phi_1$ and $\phi_2$ under NI type-II. MI continues to outperform LD in terms of bias, standard error accuracy and the expected size of estimation errors. In fact, the results are quite similar to NI type-I in that MI cuts the LD bias by about 50%. To what extent do the benefits of MI depend on the accuracy of the proxy variable? We explore this question in Figures 4 and 5.

Table 3: Bias, Overconfidence and RMSE Results for Selection on $Y$ (NI type-II)

$x, z, \varepsilon \sim N(0, 1)$, $\eta \sim N(0, 1)$, $\phi_1, \phi_2 = 1$, $n = 200$

<table>
<thead>
<tr>
<th></th>
<th>bias($\phi_1$)</th>
<th>bias($\phi_2$)</th>
<th>s.e.($\phi_1$)</th>
<th>s.e.($\phi_2$)</th>
<th>r.m.s.e.($\phi_1$)</th>
<th>r.m.s.e.($\phi_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25% Missing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listwise Deletion</td>
<td>-.108</td>
<td>-.113</td>
<td>1.00</td>
<td>1.05</td>
<td>.128</td>
<td>.134</td>
</tr>
<tr>
<td>Multiple Imputation</td>
<td>-.056</td>
<td>-.062</td>
<td>1.00</td>
<td>1.04</td>
<td>.086</td>
<td>.092</td>
</tr>
<tr>
<td>42% Missing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listwise Deletion</td>
<td>-.170</td>
<td>-.172</td>
<td>1.03</td>
<td>1.04</td>
<td>.189</td>
<td>.190</td>
</tr>
<tr>
<td>Multiple Imputation</td>
<td>-.094</td>
<td>-.094</td>
<td>1.03</td>
<td>1.03</td>
<td>.120</td>
<td>.120</td>
</tr>
</tbody>
</table>

**Note:** With selection on $Y$, 25% missing implies all values below the 25th percentile are missing. 42% missing implies all values below the 42nd percentile are missing.

Figure 4 presents the results for bias. In this figure, the first row is for the 42% MAR case, and the second row presents the results for 42% NI missingness. The first column of the figure provides the results for $\phi_1$, and the second column presents the results for $\phi_2$. What is interesting about these results is that the
Figure 4: Biases for LD and MI (42% MAR and NI)

Note: First row presents the results for 42% MAR. Second row presents the results for 42% NI missingness. First column presents the results for $\phi_1$. Second column presents the results for $\phi_2$.

performance of MI with respect to bias is bound between LD and no bias. With a non-informative proxy (i.e., pure noise), MI does as well as LD. As the signal of the proxy improves, the bias goes to zero. At the limit, the proxy is as good as the outcome. In essence, there is no missing data. Figure 5 presents the RMSE results, which are very similar to the bias results. In short, there are significant benefits to MI when one has good proxies and no cost when one has bad proxies. Clearly, our results support the conventional wisdom stated in two passages quoted by Pepinsky (2018):

Both listwise deletion and basic multiple imputation approaches can be biased under NI, in which case additional steps must be taken, or different models must be chosen, to ensure valid inferences. Thus, multiple imputation will normally be better than, and almost always not worse than, listwise deletion (King et al. 2001, 51).

the same conditions that cause multiple imputation to be severely biased also cause listwise deletion to be severely biased...Since multiple imputation is always more efficient than listwise deletion, even in this worst-case scenario it is still the preferable strategy (Lall, 2016).

4. CONCLUSION

Political methodologists have revisited recently the wisdom of multiple imputation as an alternative to listwise deletion. Multiple imputation weakly dominates listwise deletion when missingness, conditional on the observed data, is random (MAR). However, when missingness is nonignorable (NI) the decision of
Figure 5: RMSE for LD and MI (42% MAR and NI)

Note: First row presents the results for 42% MAR. Second row presents the results for 42% NI missingness. First column presents the results for $\phi_1$. Second column presents the results for $\phi_2$. 

whether to multiply impute missing data (MI) or listwise delete cases with missing data (LD) may be more complicated. In short, it helps to know what kind of missingness exists in our data. Focusing on outcomes, we propose a simple balance test to help distinguish random missingness from two types of non-ignorable missingness. Our Monte Carlo results show the balance test has good small sample properties and that MI with high quality proxies is preferable to LD.

REFERENCES


