PA

When Can Multiple Imputation Improve Regression Estimates?

Vincent Arel-Bundock[1] and Krzysztof J. Pelc[2]

[1] *Department of Political Science, Université de Montréal, Canada. Email: vincent.arel-bundock@umontreal.ca*
[2] *Department of Political Science, McGill University, Canada. Email: kj.pelc@mcgill.ca*

Abstract

Multiple imputation (MI) is often presented as an improvement over listwise deletion (LWD) for regression estimation in the presence of missing data. Against a common view, we demonstrate anew that the complete case estimator can be unbiased, even if data are not missing completely at random. As long as the analyst can control for the determinants of missingness, MI offers no benefit over LWD for bias reduction in regression analysis. We highlight the conditions under which MI is most likely to improve the accuracy and precision of regression results, and develop concrete guidelines that researchers can adopt to increase transparency and promote confidence in their results. While MI remains a useful approach in certain contexts, it is no panacea, and access to imputation software does not absolve researchers of their responsibility to know the data.

Keywords: multiple imputation, missing data, Monte Carlo simulation

Observational data in the social sciences are often incomplete. The most common approach for dealing with missing data is complete case analysis (or listwise deletion), but this strategy has important shortcomings: it ignores the valuable information carried by partially observed units, and it can introduce bias in regression coefficient estimates.

In a recent *Political Analysis* article, Lall (2016) adds to a body of work making a powerful case for an alternative: multiple imputation (MI). The author argues that listwise deletion (LWD) often introduces severe bias in regression estimates, and he applies a popular imputation routine (Honaker, King, and Blackwell 2011) to show that several published results are affected by the way analysts handle missing data.

Here, we clear up a common misunderstanding about LWD: this approach does *not* introduce bias in regression estimates, as long as the dependent variable is conditionally independent of the missingness mechanism, or when the analyst can control for the determinants of missingness.

We highlight the conditions under which MI is most likely to improve the accuracy and precision of regression results, and propose a set of best practices for empiricists dealing with missing data. The premise underlying these best practices is that while complete case analysis can be problematic, MI is no panacea: the range of circumstances under which this approach guarantees bias reduction relative to LWD is limited, and results may be sensitive to violations of the imputation model's assumptions. When results under MI and LWD diverge, analysts can make no *a priori* claim that one set of results is more credible than the other, and access to imputation software does not absolve researchers of their responsibility to know the data.[1]

Political Analysis (2018)

DOI: 10.1017/pan.2017.43

Corresponding author
Krzysztof J. Pelc

Edited by
R. Michael Alvarez

Authors' note: We thank Neal Beck, Timm Betz, Christina Davis, Tom Pepinsky, Amy Pond, and Erik Voeten for valuable comments. Replication files and supplementary materials are hosted on the Harvard Dataverse and the authors' websites. https://dataverse.harvard.edu/dataverse/pan, doi:10.7910/DVN/S9G9XS. http://arelbundock.com, https://sites.google.com/site/krzysztofpelc/.

1 In supplementary materials, we revisit one of the empirical studies replicated in Lall (2016) to illustrate the importance of the best practices we propose (Arel-Bundock and Pelc 2017). We also present results from Monte Carlo experiments designed to probe the performance of `Amelia` under different conditions.

1 When Does Listwise Deletion Introduce Bias in Regression Estimates?

After Rubin (1976), it has become standard practice to distinguish between three missingness generation mechanisms.[2] Data are said to be missing completely at random (MCAR) if the pattern of missingness is independent of both the observed and unobserved data. Data are called missing at random (MAR) if missingness depends only on observables. Data are not missing at random (NMAR) when missingness depends on unobservables.

Based on this typology, Lall (2016, 416) writes:

> "Listwise deletion is unbiased only when the restrictive MCAR assumption holds—that is, when omitting incomplete observations leaves a random sample of the data. Under MAR or [NMAR], deleting such observations produces samples that are skewed away from units with characteristics that increase their probability of having incomplete data."

This echoes King *et al.* (2001, 51), who argue that

> "inferences from analyses using listwise deletion are relatively inefficient, no matter which assumption characterizes the missingness, and they are also biased unless MCAR holds."

It is true that MI allows us to leverage more information than LWD, and that it could thus improve the efficiency of our analyses. However, the claim that LWD always introduces bias unless data are MCAR is erroneous. To demonstrate,[3] let Q_i equal 1 if the ith observation is fully observed, and 0 otherwise. A simple complete case model can be represented as:

$$\mathbf{QY} = \mathbf{QX}\boldsymbol{\beta}_c + \mathbf{Q}\boldsymbol{\varepsilon}, \quad \text{with } \mathbf{Q} = \text{diag}(Q_1, \ldots, Q_n).$$

Defining $\boldsymbol{X}_c = \mathbf{QX}$ and $\boldsymbol{Y}_c = \mathbf{QY}$, the least squares complete case estimator becomes:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_c &= (\boldsymbol{X}_c'\boldsymbol{X}_c)^{-1}\boldsymbol{X}_c'\boldsymbol{Y}_c \\ &= (\mathbf{X}'\mathbf{QX})^{-1}\mathbf{X}'\mathbf{QY} \\ &= (\mathbf{X}'\mathbf{QX})^{-1}\boldsymbol{X}'\boldsymbol{Q}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{QX})^{-1}\mathbf{X}'\mathbf{Q}\boldsymbol{\varepsilon}. \end{aligned} \tag{1}$$

Clearly, if $\boldsymbol{Q}$ is independent of $\boldsymbol{\varepsilon}$, and if the usual assumptions of the classical linear model hold, the complete case estimator is unbiased.[4] More loosely, Equation (1) shows that the OLS estimator with LWD is unbiased in the MAR cases where the pattern of missingness is unrelated to values of the dependent variable, or where we can control for the determinants of missingness.

Equation (1) also implies that complete case coefficient estimates are unbiased in the NMAR case "where the probability that a covariate is missing depends on the value of that covariate", as long as "the probability of being a complete case depends on $X_1; \ldots; X_p$ but not on Y" (Little and Rubin 2002, 43).

To be clear, the above conclusions do not depend on which variables are partially observed, but rather on the association between the values of those variables and the pattern of missingness.

2 Formal definitions can be found in many texts, including Little and Rubin (2002, 11–13).

3 We follow Jones (1996).

4 Allison (2001, fn.1) offers a more general proof: "We want to estimate $f(Y|X)$, the conditional distribution of Y given X, a vector of predictor variables. Let $A = 1$ if all variables are observed; otherwise, $A = 0$. Listwise deletion is equivalent to estimating $f(Y|X, A = 1)$. The aim is to show that this function is the same as $f(Y|X)$. From the definition of conditional probability, we have $f(Y|X, A = 1) = \frac{f(Y,X,A=1)}{f(X,A=1)} = \frac{Pr(A=1|Y,X)f(Y|X)f(X)}{Pr(A=1|X)f(X)}$. Assume that $Pr(A = 1|Y, X) = Pr(A = 1|X)$, that is, that the probability of data present on all variables does *not* depend on Y, but may depend on any variables in X. It immediately follows that $f(Y|X, A = 1) = f(Y|X)$. Note that this result applies to any regression procedure, not just linear regression."

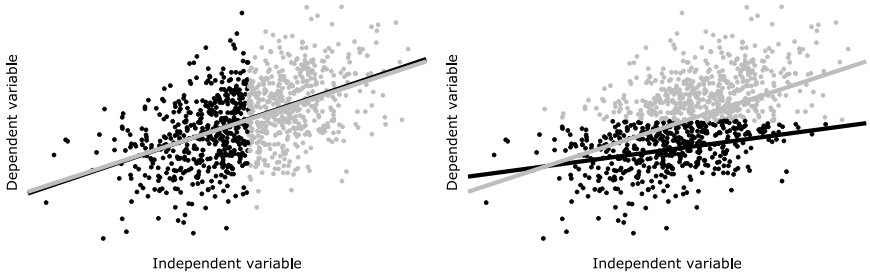

Figure 1. Linear regression under two selection mechanisms.

The outcome Y may well be unobservable for the ith individual, but as long as the reason why data are missing for that individual relates to the value of X_i and not Y_i (net of X_i), then LWD does not introduce bias in regression estimates.

These results should not be surprising to political scientists, who have long been aware of the pitfalls of "selecting cases for study on the dependent variable" (Geddes 1990). To illustrate, Figure 1 shows two simulated samples where all observed units (black) fall below an arbitrary threshold, and all unobserved units (gray) fall above that threshold.[5] The gray lines show the result of a bivariate regression model using the full data, while the black lines show analogous results based on the observed data only. In the left panel of Figure 1, sample selection is based on the values of the independent variable, and the gray and black lines overlap (no bias). In the right panel of Figure 1, sample selection is based on the values of the dependent variables, and the two linear models diverge (bias).

The practical implications are considerable. In cross-national comparisons, for instance, more complete cases are typically available for advanced democracies than for developing countries. This has led analysts to worry that their estimates may suffer from an "advanced economies" or a "pro-democracy" bias (e.g., Lall 2017, 1292).

We can distinguish between two interpretations of this problem. First, one could argue that the estimated slopes should be different in democratic and authoritarian countries, and that a full data estimate of the ("averaged") marginal effect will be sensitive to sample composition. In that case, our recommendation is that researchers model heterogeneity explicitly (Brambor, Clark, and Golder 2006; Franzese and Kam 2009), or risk misspecification bias (but not necessarily selection bias).

Second, one could think about the issue not in terms of heterogeneous marginal effects, but directly in terms of a selection problem. In that case, analysts should reflect on the nature of the association between missingness and their dependent variable. If, as in the resource curse literature, the outcome of interest is "regime type", and we suspect that this dependent variable directly affects transparency and observability (Hollyer, Rosendorff, and Vreeland 2011), then there are good reasons to worry. In contrast, when analysts can put the drivers of missingness on the right-hand side of their regression equations, LWD need not spoil the results.

2 When Can Multiple Imputation Improve Regression Estimates?

MI seems more likely to be beneficial in some contexts. First, as suggested by Equation (1), the use of LWD is largely unproblematic when data are MCAR, when missingness is solely a function of the regressors, or when control variables can purge the dependent variable of its association with the missingness generation mechanism. In those cases, MI does not reduce bias, but it could still improve efficiency.

5 X and Y are drawn from a multivariate normal with mean 0, variance 1, and covariance 0.5. The selection threshold is 0.

Second, there are good reasons to expect that MI will be most effective where missingness affects auxiliary (or control) variables, rather than the main independent or dependent variablesof interest.[6] As Little (1992, 1227) points out, if "the X's are complete and the missing values of Y are missing at random, then the incomplete cases contribute no information to the regression of Y on $X_1, \ldots, X_p$." Relatedly, White and Carlin (2010, 2928) note that "MI is likely to be beneficial for the coefficient of a relatively complete covariate when other covariates are incomplete."

Third, MI may produce better results when analysts can build an imputation model that accurately predicts the values of missing data points. When missing values are difficult to predict, or when analysts cannot leverage relevant auxiliary variables to build their prediction model, we are more likely to see large differences in coefficient estimates across imputed datasets, which would reduce the precision of the combined estimates obtained by Rubin's rules.

Fourth, an imputation routine is obviously more likely to be useful when its underlying statistical assumptions are satisfied. In particular, it is important to note that MI offers no guarantee of bias reduction unless data are MAR.[7] While we still lack systematic assessments, simple simulations have shown that LWD estimates can sometimes be *less* biased than MI estimates under NMAR (White and Carlin 2010; Pepinsky 2017).[8] MI performance can also be degraded when imputation routines make implausible distributional assumptions (e.g., multivariate normality) and data are not well-behaved.[9]

Finally, it seems reasonable to expect that MI will bring about larger improvements to precision where the proportion of fully observed units is small (White and Carlin 2010).

In sum, MI can often improve regression estimates, but this is not always the case. Because some of the assumptions that underpin LWD and MI are untestable, analysts will typically be unable to make an *a priori* claim that either set of estimates is more credible than the other. When results under LWD and MI diverge, researchers will have to exercise case-specific judgement.

3 Best Practices

To exercise this kind of case-specific judgement, researchers should take to heart the repeated admonitions of MI advocates, by developing a deep knowledge of their datasets (King *et al.* 2001; van Buuren 2012). They could also improve the credibility of their empirical work by following a set of simple best practices:

(1) Define the population of interest.
(2) Report the share of missing values for each variable and descriptive statistics for both complete and incomplete cases. Do fully observed units differ systematically from partially observed ones?
(3) Theorize the missingness mechanism. Is the pattern of missingness driven by (a) pure chance, (b) factors unrelated to the variables of interest, (c) values of the independent variables, (d) values of the dependent variable, or (e) unobservable factors? Under (a), (b), and (c), LWD can be used without fear that it will introduce bias in regression estimates.

6 In supplementary materials, we use simulations to illustrate this point.
7 Lall (2016) points out that the MAR assumption is untestable (footnote 7) and that NMAR data are ubiquitous (footnote 15).
8 Lall (2016, 5) argues that "multiple imputation is not seriously biased under [NMAR] if missingness is strongly related to observed data and thus approximates MAR (Graham, Hofer, and MacKinnon 1996; Schafer 1997; Collins, Schafer, and Kam 2001)." However, Graham, Hofer, and MacKinnon (1996) is barely germane; the simulation study in Schafer (1997, 2.5.2) is useful but perfunctory; and the main focus of Collins, Schafer, and Kam (2001) is on "[f]our conditions with different varieties of MAR missing data mechanisms." Our view is that broad pronouncements about the performance of MI under NMAR are premature, and that practitioners still lack clear guidelines to determine if their (observed) auxiliary data are rich enough for MI routines to work adequately.
9 In supplementary materials, we use simulations to illustrate how departures from multivariate normality can hinder the performance of `Amelia`, even in settings where all marginal distributions are normal. Note that other imputation procedures may relax the multivariate normality assumption, but that they typically open several other "researcher degrees of freedom." For example, the reference manual for the `mice` routine (van Buuren and Groothuis-Oudshoorn 2011) points out that the analyst needs to make *seven* main choices in the specification of the imputation model.

Under (d), MI can sometimes reduce bias, but it only offers guarantees if data are MAR and the imputation model's assumptions are satisfied. Under (e) data are NMAR and neither LWD nor MI promise unbiased estimates.

(4) Check for divergence between LWD and MI results. If estimates do diverge, which "new" observations have a strong influence on the results? Are these observations theoretically distinct?

(5) Robustness checks. Do alternative imputation procedures or tuning parameters produce different results? Does the imputation model have good predictive power? Does it fill in reasonable values for missing observations?[10]

In supplementary materials, we illustrate how these guidelines can improve statistical practice by revisiting one of the political-economy studies criticized in Lall (2016). The study we replicate meets some of the conditions listed above, and thus appears as a good *prima facie* candidate for MI. This replication exercise highlights some of the practical pitfalls of MI, and illustrates why researchers need to familiarize themselves with the data before deploying `Amelia` and concluding that MI results are more credible than LWD results.[11]

4 Conclusion

Missing data are an inevitable problem in social science. The main shortcoming of the common way of dealing with these, through LWD, is that it is done in an unthinking manner. This is where the benefit of Lall's article, and the literature to which it contributes, truly lies. We, as analysts, must show greater awareness of, and transparency about, the implications of missing data.

Unfortunately, MI is no panacea. In this note, we suggest that the range of circumstances under which this approach guarantees improvement relative to LWD is more narrow than is generally acknowledged by proponents of MI.

Taking the problem of missing data seriously means asking the type of questions raised above. Does the pattern of missingness suggest that LWD is biased, and that MI will be beneficial? What variables are truly unobserved, rather than nonexistent? Can we build an accurate prediction model to fill in missing values? And how does the expansion of the sample relate to the theory being tested? Multiple imputation requires a number of choices on the analyst's part; these must be informed by knowledge of the data and of the theory being tested.

Supplementary material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2017.43.

References

Allison, Paul D. 2001. *Missing data*, vol. 136. Thousand Oaks, CA: Sage Publications.

Arel-Bundock, Vincent, and Krzysztof Pelc. 2017. When can multiple imputation improve regression estimates? doi:10.7910/DVN/S9G9XS, Harvard Dataverse, V1.

Brambor, Thomas, William Roberts Clark, and Matt Golder. 2006. Understanding interaction models: Improving empirical analyses. *Political Analysis* 14(1):63–82.

Collins, Linda M., Joseph L. Schafer, and Chi-Ming Kam. 2001. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods* 6(4):330.

Franzese, Robert, and Cindy Kam. 2009. *Modeling and interpreting interactive hypotheses in regression analysis*. Ann Arbor, MI: University of Michigan Press.

Geddes, Barbara. 1990. How the cases you choose affect the answers you get: Selection bias in comparative politics. *Political Analysis* 2(1):131–150.

10 We concur with Graham, Hofer, and MacKinnon (1996) who write that "[b]ecause the various [imputation] procedures may be differentially sensitive to abnormalities in the data (e.g., irregularities in the minimization function, solutions near the boundary), it is always a good strategy to approach the missing data problem from different directions."

11 We show that Lall's different results are largely driven by (a) the introduction of nearly 90,000 theoretically irrelevant observations, and (b) the influence of five island nations with a combined population of about 430,000.

Graham, John W., Scott M. Hofer, and David P. MacKinnon. 1996. Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research* 31(2):197–218.

Hollyer, James R., B. Peter Rosendorff, and James Raymond Vreeland. 2011. Democracy and transparency. *Journal of Politics* 73(4):1191–1205.

Honaker, James, Gary King, and Matthew Blackwell. 2011. Amelia II: A program for missing data. *Journal of Statistical Software* 45(7):1–47, http://www.jstatsoft.org/v45/i07/.

Jones, Michael P. 1996. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association* 91(433):222–230, http://www.jstor.org/stable/2291399.

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review* 95(1):49–69.

Lall, Ranjit. 2016. How multiple imputation makes a difference. *Political Analysis* 24(4):414–433.

Lall, Ranjit. 2017. The missing dimension of the political resource curse debate. *Comparative Political Studies* 50(10):1291–1324, http://cps.sagepub.com/content/early/2016/09/06/0010414016666861.

Little, Roderick J. A. 1992. Regression with missing X's: A review. *Journal of the American Statistical Association* 87(420):1227–1237, http://www.jstor.org/stable/2290664.

Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical analysis with missing data*. Hoboken, NJ: John Wiley & Sons.

Pepinsky, Thomas. 2017. A note on listwise deletion versus multiple imputation. Working paper, Cornell University.

Rubin, Donald B. 1976. Inference and missing data. *Biometrika* 63(3):581–592, http://biomet.oxfordjournals.org/content/63/3/581.

Schafer, Joseph L. 1997. *Analysis of incomplete multivariate data*. London: Chapman & Hall.

van Buuren, Stef. 2012. *Flexible imputation of missing data*. Boca Raton, FL: CRC Press.

van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software* 45(3):1–67.

White, Ian R., and John B. Carlin. 2010. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine* 29(28):2920–2931, http://onlinelibrary.wiley.com/doi/10.1002/sim.3944/abstract.