# When Can Multiple Imputation Improve Regression Estimates?

Vincent Arel-Bundock, vincent.arel-bundock@umontreal.ca
Krzysztof J. Pelc, kj.pelc@mcgill.ca

September 7, 2017

*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.

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*a*) 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]

## 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

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 Dataverse and at the authors' websites: `https://dataverse.harvard.edu/dataverse/pan`, `http://arelbundock.com`, `https://sites.google.com/site/krzysztofpelc/`

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

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

only on observables. Data are not missing at random (NMAR) when missingness depends on unobservables.

Based on this typology, Lall (2016*a*, 3) 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 $i^{\text{th}}$ observation is fully observed, and 0 otherwise. A simple complete case model can be represented as:

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

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

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

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

[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."

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 $X_1; ...; X_p$ 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. The outcome $Y$ may well be unobservable for the $i^{\text{th}}$ 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 (grey) fall above that threshold.[5] The grey 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 1a, sample selection is based on the values of the independent variable, and the grey and black lines overlap (no bias). In 1b, 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, 2016*b*, 3).

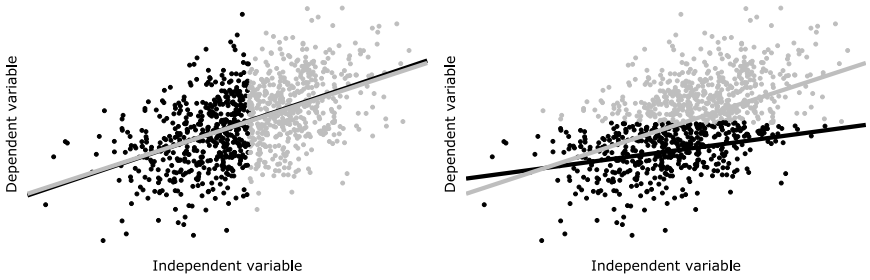
We can distinguish between two interpretations of this problem. First, one could argue that the estimated slopes should be different in democratic and authoritarian countries, and that a full data estimate of the ("averaged") marginal effect will be sensitive to sample composition. In that case, our 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 heterogenous 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

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

Figure 1: Linear regression under two selection mechanisms.

(a) Selection on the independent variable (b) Selection on the dependent variable




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.

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 variables of 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 $X_1, ..., X_p$." 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.

[6] In supplementary materials, we use simulations to illustrate this point.

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, 2016).[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

[7] Lall (2016*a*) points out that the MAR assumption is untestable (footnote 7) and that NMAR data are ubiquitous (p. 5).

[8] Lall (2016*a*, 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.

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. 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*a*). 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 listwise deletion, 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.

[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.

Unfortunately, multiple imputation is no panacea. In this note, we suggest that the range of circumstances under which this approach guarantees improvement relative to listwise deletion 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 listwise deletion is biased, and that multiple imputation will be beneficial? What variables are truly unobserved, rather than non-existent? 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.

## 5 SUPPLEMENTARY MATERIALS: REPLICATION

To illustrate the benefits of the guidelines proposed in text, we replicate one of the political economy studies criticized in Lall (2016*a*).[12] One of the key takeaways from our note and from the literature on MI is that proper use of this technique requires serious consideration and a deep understanding of the data at hand. Unfortunately, this means that carefully reviewing all forty-two of the studies replicated by Lall is not feasible. Instead, we focus on one study, Pelc (2011), which we are familiar with, and which appears as a good *prima facie* candidate for imputation.

### *5.1 Argument*

Pelc (2011) asks what explains variation in the level of flexibility that countries inject into their trade tariffs. Such flexibility, called "binding overhang," is equivalent to the difference between the trade duties actually levied at the border and the "bound," or maximum rate, that countries commit to, and which they cannot legally exceed. The greater the difference, the greater countries' flexibility to legally raise trade protection. Pelc argues that binding overhang generates costly uncertainty, and shows that governments that enjoy alternative sources of flexibility—floating exchange rates or the ability to use antidumping measures—retain lower levels of binding overhang.[13] Lall (2016*a*) applies multiple imputation to Pelc's replication dataset, and produces new coefficient estimates that fail to cross standard thresholds of statistical significance.

### *5.2 Descriptive statistics*

Table 1 shows descriptive statistics from Pelc's replication dataset. Most missing observations are found in the dependent variable *Binding Overhang*, and three of the independent variables: *Products Imports*, *Floating Currency*, and *Regime Type*. Interestingly, the sample of completely observed units is descriptively similar to the sample of incompletely observed units: most of the variables' means are similar in the complete and incomplete data (Table 1). But if there are few differences in descriptive statistics across complete and incomplete rows of the dataset, there are important differences in the variables' means between units for which the dependent variable is observed or missing. In particular, the proportion of *Floating Exchange Rates* – the key independent variable – is 10% for units where the dependent variable is observed, but much smaller where the dependent variable is missing (2%). This

[12] For full replication materials for these estimations and the simulations below, see Dataverse doi:10.7910/DVN/S9G9XS (Arel-Bundock and Pelc, 2017).

[13] Since countries' bound duties rarely move over time, data are cross-sectional. Observations are made at the country-product level during a country's first year following WTO accession.

suggests that the patterns of missingness on the left and right-hand sides of Pelc's regression equation may be conceptually different.

*Missingness in the dependent variable.* As we noted above, a large fraction of observations (23%) in Pelc's replication data do not carry information about the dependent variable, *Binding Overhang*. Indeed, about 40% of the "new" observations that Lall's imputation introduces originally showed missing values on that key variable. But just because we *can* impute those missing values does not mean that we *should*. Sometimes, knowledge of the data dictates that an observation remain "missing."

Recall the meaning of *Binding Overhang*: it reflects the flexibility of a country's tariff rate commitments at the WTO. A country's international commitment on a given tariff line only becomes binding once the WTO records it; the organization holds a complete record of all binding tariff commitments.[14] By construction, this variable is complete.

"Missingness," in this case, reflects the fact that some countries choose not to make commitments on every product. For instance, Bangladesh has long been unwilling to bind itself within the international trade regime, refusing to commit to a cap on a great number of its tariff lines. For each of the 696 products where Bangladesh remains unbound to this day, the replication dataset from Pelc (2011) records a missing value for the dependent variable. With imputation, these blank spaces get filled in, creating out of whole cloth a total of $89,030$ tariff commitments that never took place.

These are not "unobserved" data points; they are "non-existent."[15] Retaining such missing observations in a dataset is inconsequential when applying listwise deletion. But when using multiple imputation, care must be taken to remove theoretically irrelevant observations before regression analysis.

*Missingness in the independent variables.* Turning to right-hand side variables, it is interesting to note that most of the missing data in Pelc are found in *Products Im-*

[14] When the WTO itself calculates a country's official average tariff rate, it thus excludes unbound tariff lines. WTO Statistics. `https://www.wto.org/english/res_e/statis_e/popup_indicator_help_e.htm`.

[15] One interesting possibility is that the original results in Pelc (2011) could suffer from a form of selection problem not highlighted by Lall. We could think of tariff negotiations as a two-step process. First, countries decide whether they want to be bound on a given tariff line. Second, they choose the specific value of the tariff they want to commit to. This problem lies outside the scope of the replication, but it is worth noting that the share of country-product observations with floating exchange rates is much higher in the sample where tariff commitments are made than where countries refuse to be bound (10% vs. 2%, respectively - Table 1). If we think of unbound tariff lines as having very high flexibility, introducing the "dogs that didn't bark" into the estimation amounts to including many new observations with fixed exchange rates *and* very high binding overhang. In this view, Pelc's original estimates appear more conservative.

*ports*, *Floating Currency*, and *Regime Type*. In a series of tests not reported here, we found that the observations dropped by LWD due to missingness in the first two variables generally conform to Pelc's expectations; including those units in the sample via MI or other means does not affect the results.[16] We thus focus our attention on the *Regime Type* variable, whose effect on sample composition turns out to be consequential.

In the original estimation, over 36,000 observations were excluded by listwise deletion due to lack of information about democracy. Does ignoring those observations introduce bias in the estimates? We can offer a preliminary answer to this question and develop some intuition about the threat to inference by explicitly theorizing the missingness generation mechanism.

### *5.3 Why are data missing? Should we expect listwise deletion to bias the regression estimates?*

Recall that the complete case OLS estimator is biased when the dependent variable remains associated with missingness after we condition on the regressors. This can happen if (a) values of the dependent variable directly determine if a given case is fully observed, or (b) some unobserved variable drives both values of the dependent variable *and* the pattern of missingness.

*A priori*, it seems unlikely that the level of tariff commitment flexibility directly affects whether we can measure a country's level of democracy. The question thus becomes: Can we think of (unobserved) variables which drive both values of the dependent variable *and* the pattern of missingness? The answer is "yes."

Case selection for the *Regime Type* variable is a deterministic function of population size: the Polity Project only includes countries with a population greater than $500,000$.[17] Similar sample selection strategies are often adopted by international organizations and scholars who choose not to collect data on extremely small countries, island nations, and semi-autonomous territories.

There are good reasons why political economy theories may apply differently to such anomalous units. Consider five of the small countries excluded from Pelc's original analysis because they do not feature in Polity IV: Antigua and Barbuda, Dominica, Grenada, St-Kitts and Nevis, and St-Lucia. Polity excludes such countries in part because in spite of their formal sovereign status, states with populations of less than 50,000 (as is the case of St-Kitts in the sample period) give rise to different

[16] We manually included information on *Floating Currency* for ten countries/regions which were not coded in the currency regime dataset of Reinhart and Rogoff. We use two alternative approaches to include the observations with missing *Products Imports*: We drop that variable from the regression model and use listwise deletion, or we apply multiple imputation to the subset of observations which show missing *Products Imports* but are otherwise complete.

[17] Polity IV Dataset User's Manual. 2015. `www.systemicpeace.org/inscr/p4manualv2015.pdf`

expectations over the presence of political institutions. One can argue that similar concerns arise in the case of trade: these island nations are linked by their use of the East Caribbean dollar, which has been pegged to the US dollar since 1976. They are also highly dependent on imports, so they protect only a small number of domestic industries, and their tariff line commitments thus feature relatively low binding overhang. In this case, the pattern of missingness is clearly related to values of the dependent variable; taking those five countries into account pulls the estimates in a direction counter to the theory's expectations.[18]

This does not pose a problem as such: Pelc's argument is stated in probabilistic terms, and researchers rarely expect every unit to behave as theory predicts. However, as is well known, the leverage of an observation in regression analysis is a function of its proximity to the centroid of the data (Greene, 2011, 99-100). Since island nations tend to be exceptional along most dimensions, they can exert an influence on the overall results which is disproportionate to their importance in the world economy. In other words, by including countries that are purposefully left uncoded by Polity or the World Bank, multiple imputation risks introducing outliers, with attendant consequences on the overall results.

### *5.4 Why do results diverge under MI and LWD?*

Table 2 shows how each of the choices described above affects Pelc's results. The first two columns show, respectively, the original estimation and the replication using multiple imputation to fill in missing values in all rows of the dataset. In Model 3, we exclude the theoretically irrelevant observations with missing dependent variable.[19] In Model 4, we also exclude the five small East Caribbean states mentioned above.[20]

These two corrections suffice to restore the original results that Lall claims as "disappeared."[21] Lall's divergent results thus appear driven by the introduction of

[18] Note that the European Union, for example, counts as a single unit of observation in the analysis, on the same order as each island nation that Lall introduces through imputation. Given that EU member states are part of a customs union, this is the correct treatment, but this comparison highlights how MI may give disproportionate weight to units that occupy a marginal position in the world economy.

[19] We also fill in true values of the main independent variables in a few cases where data were originally missing, but for which reliable information are now available. This yields new observations for 12 countries on the exchange rate variable, and 2 countries on trade remedies usage. These countries are missing in the Reinhart and Rogoff dataset because of uncertainty about the precise exchange rate regime (e.g. whether a country's de facto regime corresponds to a moving band of $+/-$ 2% or $+/-$ 5%), but filling these in for the binary variable "Fully Floating Currency" proves straightforward: all but the EU are not fully floating for 1995, or the relevant year of WTO entry. The other countries are Taiwan, Angola, Namibia, Fiji, Oman, Cuba, Macao, Macedonia, St-Kitts, Rwanda, and Djibouti. As for the trade remedy variable, the Bown (2011) data lack observations for Trinidad and Tobago and St-Kitts and Nevis, neither of which was a trade remedy user at the moment of its WTO accession.

[20] The results are very similar if we drop all nine of the very small countries that were originally unaccounted for by the Polity Project.

[21] Lall replicates three other models from Pelc (2011). The corrections we propose here also bring

nearly 90,000 non-existent country commitments and, less problematically, by the addition of observations for five outliers: Antigua and Barbuda, Dominica, Grenada, St-Kitts and Nevis, and St-Lucia.

What are we to make of this? Model 3 should be uncontroversial, since it corrects a substantive mistake. However, since Pelc (2011) does not explicitly consider whether extremely small countries are subject to different incentives from others, discussion of Model 4 necessarily amounts to *post hoc* theorizing about scope conditions. The question then becomes whether the weight of these five island nations, with a combined population of less than 450,000, should serve as falsifying evidence against a relationship which holds for the rest of the sample.

### *5.5 Robustness checks*

In our paper, we recommended that applied researchers probe the robustness of their results by comparing results using different imputation procedures and tuning parameters. In that respect, Lall (2016*a*) should be commended, since he replicated a random sample of C/IPE studies using both the `Amelia` and `mice` imputation routines. In the case at hand, the choice of imputation routine turns out to be much less consequential than the problems we discussed above.

Yet it is worth noting that imputing Pelc's replication data using `Amelia`'s default settings (as Lall does) produces highly implausible values for many variables. Consider the imputed values of *Binding Overhang* for the aforementioned case of Bangladesh. There, Lall's "complete" datasets show (imputed) commitments far lower than the country's (actual) average tariff, moving the average Bangladeshi bound rate from 167% in the original data to 80% in the imputed data. But, if anything, unbound tariff lines should be thought of as having maximum rates far *higher* than the average, since they can be raised at will. `Amelia` also produces a range of impossible values, such as negative bound tariffs of -107%—effectively a commitment by countries to pay exporters the full value of their exports at the border.

Referring back to our proposed best practices, empiricists should let their knowledge of the data guide their assessment of imputed data. When these appear implausible, it may be that MI is being used in ways that the data-generating process may not support.

those models in line with the original published estimates, with Models 1, 3, and 4 from Pelc's Table 2 yielding "conclusive" results, and Model 2 producing "mixed" results.

Table 1: Descriptive statistics

| | | | | Mean | | | | |
|---|---|---|---|---|---|---|---|---|
| | Missing % | Range | SD | Full | Complete | Incomplete | Obs.DV | Miss.DV |
| Binding Overhang | 23.10 | 864.63 | 26.90 | 18.63 | 18.67 | 18.59 | 18.63 | |
| Logged Products Imports | 24.50 | 13.82 | 2.32 | 3.29 | 2.98 | 3.54 | 3.37 | 3.07 |
| Fully Floating Currency | 13.60 | 1.00 | 0.27 | 0.08 | 0.10 | 0.06 | 0.10 | 0.02 |
| Regime | 9.40 | 108.00 | 10.24 | 3.29 | 3.19 | 3.41 | 4.19 | 0.27 |
| Logged GDP | 3.80 | 10.39 | 2.52 | 23.97 | 24.12 | 23.78 | 24.16 | 23.27 |
| MFN | 2.90 | 871.40 | 13.67 | 12.08 | 11.46 | 12.89 | 11.44 | 14.53 |
| Logged GDP per capita | 2.50 | 5.66 | 1.55 | 7.83 | 8.00 | 7.62 | 8.02 | 7.20 |
| Remedies User | 1.60 | 1.00 | 0.47 | 0.34 | 0.36 | 0.30 | 0.38 | 0.21 |
| Agricultural Product | 0.00 | 1.00 | 0.32 | 0.11 | 0.09 | 0.15 | 0.14 | 0.03 |
| LDC | 0.00 | 1.00 | 0.36 | 0.15 | 0.19 | 0.10 | 0.10 | 0.32 |
| Recent Entrant | 0.00 | 9.00 | 1.84 | 0.69 | 0.57 | 0.86 | 0.76 | 0.49 |

Table 2: The Effect of Policy Substitutes on Tariff Flexibility: Replicated vs. Imputed Data

| | (1) Original Estimation | (2) Lall Replication | (3) Tariffs Correction | (4) Carribean Correction |
|---|---|---|---|---|
| Applied Rate | -0.522 | -0.348 | -0.403 | -0.441 |
| | (0.081) | (0.064) | (0.081) | (0.083) |
| Logged GDP per capita | -1.708 | -1.243 | -0.682 | -1.719 |
| | (3.300) | (1.704) | (2.278) | (2.329) |
| Logged GDP | 0.894 | -2.850 | -2.159 | -0.664 |
| | (1.315) | (0.900) | (1.054) | (1.040) |
| Regime | 0.125 | 0.177 | 0.188 | 0.173 |
| | (0.319) | (0.131) | (0.217) | (0.234) |
| Logged Products Imports | 0.091 | 0.032 | 0.006 | -0.057 |
| | (0.165) | (0.112) | (0.128) | (0.129) |
| LDC dummy | 4.667 | 3.407 | 5.148 | 6.989 |
| | (7.453) | (4.754) | (6.702) | (6.409) |
| Agricultural Product | 23.929 | 19.042 | 18.816 | 17.446 |
| | (3.773) | (2.800) | (3.046) | (3.061) |
| Recent Entrant | -5.023 | -4.612 | -4.577 | -4.602 |
| | (0.631) | (0.448) | (0.617) | (0.680) |
| Fully Floating Currency | -9.909 | -3.965 | -7.758 | -10.356 |
| | (3.828) | (3.856) | (4.657) | (4.581) |
| Remedies User | -16.168 | -8.696 | -11.084 | -12.866 |
| | (7.718) | (5.360) | (6.678) | (6.587) |
| Constant | 22.052 | 104.739 | 85.488 | 58.503 |
| | (38.362) | (25.137) | (31.422) | (30.859) |
| N | 163097 | 385798 | 296768 | 284286 |

Dependent variable is binding overhang. OLS estimates with robust standard errors in parenthesis clustered on common country. Column (1) is the original regression from Pelc 2011. Column (2) is Lall's replication after imputation of all missing observations. Column (3) is model (2), excluding non-existent tariffs, and adding known currency regime data and trade remedies data. Column (4) is model (3), excluding imputed data for five Caribbean countries that Polity IV avoids coding due to their small size.

## 6 SUPPLEMENTARY MATERIALS: SIMULATION

In the text, we developed rules of thumb to help researchers identify the conditions under which MI is most likely to be beneficial. Here, we use a set of very simple Monte Carlo experiments to illustrate.

### *6.1 New assumptions, new problems?*

`Amelia` combines an expectation-maximization algorithm with a bootstrap approach to impute missing values in partially-observed datasets. It makes two main assumptions: data must be MAR and be distributed following a multivariate normal law (Honaker, King and Blackwell, 2011, 3-4).

Proponents of MI often claim that the approach performs well under NMAR. Unfortunately, as we mentioned in Footnote 8 of the text, evidence in support of that contention is scant. One important problem is that the NMAR concept covers a vast array of potential dependence patterns, and that any performance assessment will be highly dependent on the specific data generating process under investigation. This means that any attempt to compare the performance of LWD and MI in NMAR data will at best be partial in scope. That said, a recent unpublished manuscript by Pepinsky (2016) raises some concerns. Based on extensive Monte Carlo simulations, the author concludes that "multiple imputation yields results that are frequently more biased than listwise deletion when data are [NMAR] [...] even with very strong correlations between fully observed variables and variables with missing values, such that the data are very nearly MAR." In short, while we do not have access to strong evidence either way, there are good reasons to remain cautious in the (quite typical) case where researchers are unable to claim that the MAR assumption holds.

The multivariate normality assumption also raises potential issues since, as `Amelia`'s authors concede, it is "often a crude approximation to the true distribution of the data" (Honaker, King and Blackwell, 2011, 4).[22] And even if analysts can use truncation or transformations to make individual variables look more "normal", the multivariate normal assumption imposes requirements beyond marginal distributions: it also constrains the *structure of relationships* between variables. Below, we show that even if every variable, on its own, is standard normal, `Amelia`'s performance can be severely degraded when the *dependence* between variables is not normal.

### *6.2 Simulation design*

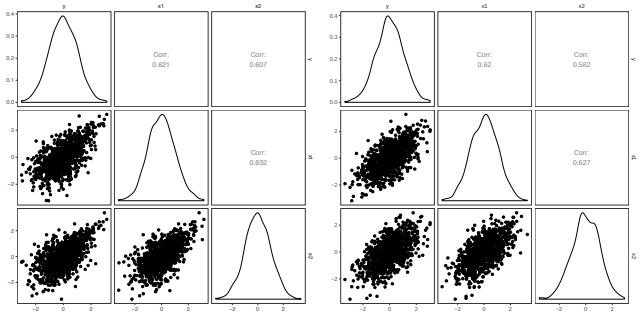
We wish to use a linear regression model to estimate the association between regressand $y$ and regressor $x_1$, controlling for $x_2$. To study the effect of deviations from

[22] Other imputation procedures relax the multivariate normality assumption, but open several "researcher degrees of freedom." For example, the `mice` routine (van Buuren and Groothuis-Oudshoorn, 2011) requires that the analyst makes seven choices in the specification of the imputation model.

multivariate normality, we draw values for $y, x_1, x_2$ using four different random numbers generators ($N = 1000$).[23] The first produces multivariate normal data with mean zero, variance one, and covariances equal to $1/3$. The other three use *Clayton*, *Gumbel*, and *Frank* copulas to produce data whose marginal distributions are standard normal, but whose dependence structure is non-normal.[24] Figure 2 illustrates the difference between variables drawn from a Gumbel copula and others drawn from a multivariate normal. In both cases, the marginal distributions (density plots on the diagonals) are standard normal. However, the scatterplots show that the structure of dependence is slightly different in the Gumbel and normal data.

Figure 2: Variables with normal marginal distributions drawn from three Archimedean copulas and a multivariate normal (N=1,000).

(a) Gumbel (b) Multivariate Normal




To see how the missingness mechanism affects the performance of MI, we eliminate observations in each dataset by drawing binary indicator $Q_i$ from a binomial distribution, where the probability that unit $i$ is completely observed is given by:

$$Pr(Q_i = 1) = logit(\theta_1 + \theta_{x_1} x_{1i} + \theta_{x_2} x_{2i} + \theta_y y_i). \tag{2}$$

$\theta_1$ controls the share of partially observed cases. When $\theta_{x_1} = \theta_{x_2} = \theta_y = 0$, data are MCAR and we expect no bias. When $\theta_{x_1} = \theta_{x_2} = 1$ and $\theta_y = 0$, missingness is a function of the observed regressors (MAR), but it is conditionally independent of the regressand; again, we expect no bias. When $\theta_y = 1$ and $\theta_{x_1} = \theta_{x_2} = 0$,

[23] In a series of tests not reported here, we found that adding correlated auxiliary variables to the imputation stage does not materially affect our overall conclusions for the multivariate normal case. Unfortunately, since we cannot manipulate individual covariance parameters with the `copula` package for `R`, it is impossible to use imputation-only variables in the other experiments.

[24] Copulas are multivariate probability distributions for which all marginals are uniform over [0,1]. This property allows us to use the probability integral transformation to model the dependence structure between variables separately from their marginal distributions (Yan et al., 2007). The (arbitrary) tuning parameters that control the strength of association between $x_1$, $x_2$, and $y$ are 1.8 (*Gumbel*), 5.4 (*Frank*), and 1.5 (*Clayton*).

the outcome variable remains associated with the pattern of missingness, even after we control for $x_1$ and $x_2$. In practice, this situation could arise when the outcome variable itself drives missingness, or when an unobserved variable determines both the outcome and missingness. In such cases, we expect LWD to introduce bias in regression estimates.

To assess the performance of MI when different variables need to be imputed, we use the $Q_i$ indicator to create three versions of each partially-observed dataset. To begin, we use Equation 2 to erase values of the dependent variable, but leave the $x_1$ and $x_2$ regressors intact. Then, we repeat the exercise with the other two independent variables. All datasets are imputed ten times using the `Amelia` software.

Figure 3 shows the mean absolute deviation from full-data estimates of the $x_1$ coefficient under different data generation mechanisms. Four main conclusions emerge.

First, columns 3 and 4 show that MI does not materially improve upon LWD when data are MCAR or where we can control for the determinants of missingness.[25] This is consistent with the analytical results we presented in text, which show that LWD estimates are unbiased when missingness is conditionally independent of the dependent variable.

Second, when the missingness mechanism is related to the dependent variable and data are multivariate normal, imputing data with `Amelia` can yield important benefits.

Third, improvements with MI seem particularly large when the control variable ($x_2$) is affected, rather than the main independent or dependent variables of interest. This makes sense because, 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 $X_1, ..., X_p$."[26] Conversely, the imputation of auxiliary variables yields a benefit because it allows for estimation based on units that are otherwise fully observed with respect to the $Y$ and the $X$'s of interest.

Fourth, the rest of Figure 3 shows that even if all the variables in our experiments are standard normal, deviations from *multivariate* normality can severely degrade the imputation algorithm's performance. Indeed, as we manipulate the structure of dependence between variables, we see that LWD often *outperforms* `Amelia`.

This is not to say that all departures from multivariate normality will hinder the

[25] `Amelia` does seem to produce slight improvements at very high levels of missingness (e.g., 85%). However, this is only the case when the control variable is missing, and not when either the outcome or the main independent variables are missing. Note that these simulations probably understate the benefits of MI where analysts can leverage auxiliary variables with high predictive power.

[26] Leaving units with missing outcome in the dataset for imputation can still help estimation if it improves the imputation model for missing values of the regressors.

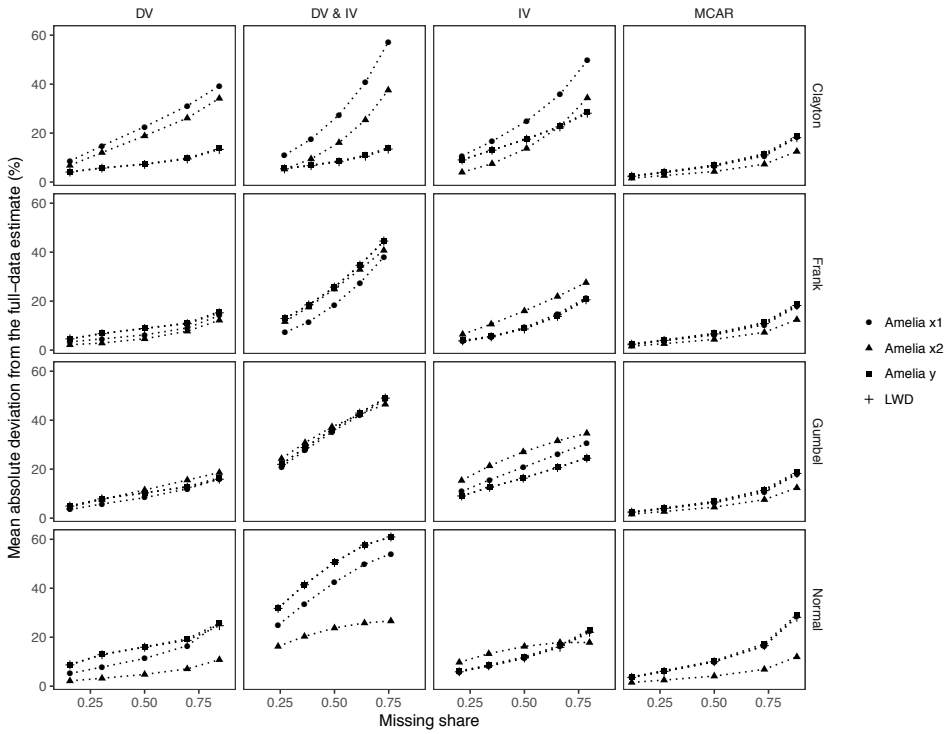
imputation procedure. The copulas we used above were chosen for convenience[27], and because they are widely used in the statistical literature. We do not argue that these distributions represent better approximations of social phenomena, nor do we claim that all non-normal multivariate data will degrade the performance of `Amelia`. Nevertheless, the results in Figure 3 are interesting because, to our eyes at least, the *Clayton*, *Frank*, and *Gumbel* data do not look more "atypical" than the observational data we regularly work with. If seemingly innocuous departures from multivariate normality can have such deleterious effects on `Amelia`'s performance, one wonders how well it can be expected to work in real-life settings, where data are messier and the multivariate central limit theorem does not rule all.

In sum, Monte Carlo experiments support our earlier contention that missingness does not substantially impair our ability to obtain accurate regression estimates using LWD, as long as we can control for the determinants of missingness. MI using `Amelia` can sometimes improve our estimates, but the procedure relies on two very strong assumptions: MAR and multivariate normality. In practical applications, NMAR data are ubiquitous, the MAR condition is untestable, and existing evidence does not allow us to conclude that MI dominates LWD when MAR is violated. Multivariate normality is often a poor descriptor for real-world data, and departures from that dependence structure can severely degrade the performance of the imputation model.

This discussion suggests that analysts would do well to probe the sensitivity of their results by trying different MI routines and tuning parameters. Moreover, when results under MI and LWD diverge, analysts will generally be unable to make an *a priori* claim that one set of results is more credible than the other. Case-specific judgment and knowledge of the data remain important.

[27] Random number generators are readily available for `R` (Hofert et al., 2016).

Figure 3: Performance of different estimation procedures in 5,000 Monte Carlo simulations.



## REFERENCES


Allison, Paul D. 2001. *Missing data*. Vol. 136 Sage publications.

Arel-Bundock, Vincent and Krzysztof J. Pelc. 2017. "When Can Multiple Imputation Improve Regression Estimates?".
**URL:** *http://dx.doi.org/10.7910/DVN/S9G9XS*

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*. 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.

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.

Greene, William H. 2011. "Econometric Analysis, Fifth Edition.".

Hofert, Marius, Ivan Kojadinovic, Martin Maechler and Jun Yan. 2016. *copula: Multivariate Dependence with Copulas*. R package version 0.999-15.
**URL:** *http://CRAN.R-project.org/package=copula*

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.
**URL:** *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.
**URL:** *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*a*. "How Multiple Imputation Makes a Difference." *Political Analysis* p. mpw020.

Lall, Ranjit. 2016*b*. "The Missing Dimension of the Political Resource Curse Debate." *Comparative Political Studies* p. 0010414016666861.
**URL:** *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.
**URL:** *http://www.jstor.org/stable/2290664*

Little, Roderick JA and Donald B. Rubin. 2002. *Statistical analysis with missing data*. John Wiley & Sons.

Pelc, Krzysztof J. 2011. "How States Ration Flexibility: Tariffs, Remedies, and Exchange Rates as Policy Substitutes." *World Politics* 63(4):618–646.

Pepinsky, Thomas. 2016. "A Note on Listwise Deletion Versus Multiple Imputation.".

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

Schafer, Joseph L. 1997. *Analysis of incomplete multivariate data*. CRC press.

van Buuren, Stef. 2012. *Flexible imputation of missing data*. CRC press.

van Buuren, Stef and Karin Groothuis-Oudshoorn. 2011. "mice: Multivariate imputation by chained equations in R." *Journal of statistical software* 45(3).

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.
**URL:** *http://onlinelibrary.wiley.com/doi/10.1002/sim.3944/abstract*

Yan, Jun et al. 2007. "Enjoy the joy of copulas: with a package copula." *Journal of Statistical Software* 21(4):1–21.