# Listwise Deletion in High Dimensions

J. Sophia Wang[1] and P. M. Aronow[2]

[1] Graduate Student, Department of Political Science, Yale University, New Haven, CT, USA. E-mail: jinghong.wang@yale.edu
[2] Associate Professor, Departments of Political Science, Biostatistics, and Statistics and Data Science, Yale University, New Haven, CT, USA. E-mail: peter.aronow@yale.edu

## Abstract

We consider the properties of listwise deletion when both *n* and the number of variables grow large. We show that when (i) all data have some idiosyncratic missingness and (ii) the number of variables grows superlogarithmically in *n*, then, for large *n*, listwise deletion will drop all rows with probability 1. Using two canonical datasets from the study of comparative politics and international relations, we provide numerical illustration that these problems may emerge in real-world settings. These results suggest that, in practice, using listwise deletion may mean using few of the variables available to the researcher.

*Keywords:* missing data, listwise deletion, high dimensional inference

## 1 Introduction

Listwise deletion is a commonly used approach for addressing missing data that entail excluding any observations that have missing data for any variable used in an analysis. It constitutes the default behavior for standard data analyses in popular software packages: for example, rows with any missing data are by default omitted by the *lm* function in R (R Core Team 2020), the *regress* command in Stata (Stata.com 2020), and the *glmnet* function in the R package of the same name (Friedman, Hastie, and Tibshirani 2010).

However, scholars have increasingly recognized that listwise deletion may not be a generally appropriate research method to handle missing data. While a common critique focuses on the plausibility of the "missing completely at random" assumption (Allison 2001, 6–7; Cameron and Trivedi 2005, 928; Little and Rubin 2019, 15; Schafer 1997, 23), issues about efficiency in estimation have also been raised (Allison 2001, 6; Berk 1983, 540; Schafer 1997, 38). Namely, since listwise deletion discards data, the resulting estimators can be inefficient relative to approaches that use more of the data (e.g., imputation methods).

These issues have been raised to an audience of political scientists (Honaker and King 2010; King *et al.* 2001; Lall 2016), but the manner in which listwise deletion can hinder the researcher has been underappreciated. Namely, if the researcher seeks to use many variables with missingness, it may be impossible altogether to draw any statistical conclusion whatsoever. Accordingly, the use of listwise deletion may imply a severe restriction on variables used in an analysis.

The primary purpose of this note is to make this argument rigorous by considering the properties of listwise deletion when *both* the number of variables *k* and the number of units *n* are large. We show that when (i) all variables have some idiosyncratic missingness and (ii) the number of variables grows with *n* at any superlogarithmic rate, listwise deletion will yield no usable data asymptotically with probability 1. In Supplementary Material A, we report numerical illustrations to shed light on finite-*n* properties under our assumptions.

We then demonstrate real-world implications by considering two real-world datasets: the Quality of Government (QoG) dataset (Teorell *et al.* 2021) and the State Failure dataset (King and Zeng 2001, 2007). We first report on the empirical patterns of missingness in these datasets. We then conduct a simulation study by randomly subsampling from the variables in these datasets. We show that, even when a qualitatively small number of variables have been chosen from these

***Political Analysis* (2023)**
**vol. 31: 149–155**
**DOI:** 10.1017/pan.2022.5

**Published**
2 March 2022

**Corresponding author**
P. M. Aronow

**Edited by**
Jeff Gill

datasets, very little of the data may remain after listwise deletion. Taken together, we conclude that listwise deletion is simply not viable in many data-analytic settings.

## 2 Theory

We consider a fairly general setting designed to accommodate probabilistic missingness in data. Our results will apply to any estimator, algorithm, or procedure (including, e.g., variable selection or regression) on datasets in this setting, so long as this researcher's chosen procedure depends on listwise deletion.

Before we proceed, we introduce some notation. Let $n$ be the number of observations. Let $k$ be the number of variables (columns) in the dataset. Let $M_{ij}$ be a random indicator variable for whether or not the $j$th variable in the $i$th row is missing. For notational convenience, we will let $\mathbf{M}_{ij}$ represent the random vector collecting the missingness indicators up to variable $j$ $(M_{i1}, M_{i2}, \ldots, M_{ij})$.

We will invoke three key assumptions for our results. These assumptions are general with respect to standard assumptions about missingness—that is, they are compatible with both missing-at-random and missing-not-at-random data generating processes. The first of these assumptions is mutual independence of missingness across rows. This would be violated when, for example, observations are clustered, as would often be the case for longitudinal data.

Assumption 1. *All rows of the data* $((M_{11}, \ldots, M_{1k}), \ldots, (M_{n1}, \ldots, M_{nk}))$ *are mutually independent.*

We also assume that there is some idiosyncratic missingness in each variable. Namely, we will assume that there is a factorization of the data such that all conditional probabilities that an observation is missing are bounded away from zero, given whenever prior variables are observed. Substantively, this assumption rules out the case where, for example, one variable is always observed whenever another variable is observed.

Assumption 2. *There exists a* $q_* \in [0, 1)$ *such that for all* $i$,

- $\Pr(M_{i1} = 0) \leq q_*$ *and*
- $\Pr(M_{ij} = 0 | \mathbf{M}_{i(j-1)} = 0) \leq q_*$, *for all* $j \in \{2, \ldots, k\}$ *such that* $\Pr(\mathbf{M}_{i(j-1)} = 0) > 0$.

Note that this is an assumption about the presence of missingness and not about how such missingness relates to outcomes. Thus, Assumption 2 is compatible with both missing-at-random and missing-not-at-random data generating processes.

With some additional notation, Assumption 2 can be weakened to only require the existence of an index ordering over $j$ such that the condition holds. We also note that Assumption 2 may be unrealistic in settings where missingness is always identical across some groups of variables (e.g., because two or more variables come from a common data source). In such settings, our results can be generalized to the setting where $k$ refers to the number of groups of variables, rather than the number of variables themselves. We formalize this extension in Supplementary Material B.

### 2.1 Results

We begin by application of elementary probability theory to yield the following finite-$n$ result. In words, this result establishes an exact lower bound on the probability that all rows of the data will suffer from some missingness. In such instances, listwise deletion would yield no usable data.

Lemma 3. *Under Assumptions 1 and 2, the probability that listwise deletion removes all rows is* $p_{all} \geq (1 - q_*^k)^n$.

This result will be helpful in proving our main result shortly in Proposition 5. We can now consider the asymptotic properties of listwise deletion, letting both $k$ and $n$ tend to infinity. To

https://doi.org/10.1017/pan.2022.5 Published online by Cambridge University Press

do so, we embed the above problem into a sequence. We let $k_n = f(n)$, where $f$ has range over the natural numbers, and allow $M_{ij,n}$ and therefore $q_{ij,n}$ to vary at each $n$. To ease exposition, we omit notational dependence on $n$.

We have our third and final assumption: $k$ grows superlogarithmically in $n$. This is the primary point of divergence from standard (low-dimensional) theoretical treatments, under which $k$ is assumed to be fixed regardless of $n$. We emphasize here that Assumption 4—like other assumptions about asymptotics—needs not be thought of as a literal growth process that a researcher might follow, but rather an approximation of probabilistic behavior when $n$ (and here, also $k$) are large.[1]

Superlogarithmic rates can be extremely slow, and include any polynomial rate of growth (e.g., $n^c$ for any $c > 0$). Thus, our results can speak to cases where $n$ is large, $k$ is large, but $k \ll n$. To see how slow these rates can be, our results would include the rate $\lfloor \log(n)^{1.1} \rfloor$, which would permit use of 2 variables with 10 observations, 8 variables with a thousand observations, and 17 variables with a million observations. This assumption encompasses rates that are slow enough that they would not normally preclude good asymptotic behavior for most standard estimators. For example, see the mild assumptions invoked by Lai, Robbins, and Wei (1978) for convergence of the least squares estimator.

Assumption 4. *The number of covariates grows superlogarithmically in n, so that* $\lim_{n\to\infty} \frac{f(n)}{log(n)} = \infty$.

Assumption 4 can be equivalently written in asymptotic shorthand notation as $k = \omega(\log n)$. The following proposition demonstrates that when Assumptions 1, 2, and 4 hold, then the probability of listwise deletion yielding no usable data tends to 1 as $n \to \infty$.

Proposition 5. *Under Assumptions 1, 2, and 4,* $\lim_{n\to\infty} p_{all} = 1$.

Thus, we have shown that even modest rates of growth in the number of covariates can render any resulting statistical inference asymptotically impossible with listwise deletion. Our results, however, critically depend on the assumption that the number of covariates exhibits such growth in $n$; otherwise, it is possible that $\lim_{n\to\infty} p_{all} = 0$.

Our results are supported by numerical illustrations in Supplementary Material A, which also demonstrate finite-$n$ implications. These results demonstrate that our theoretical results are most relevant in finite-$n$ settings when rates of idiosyncratic missingness are high. When there are low rates of idiosyncratic missingness (e.g., 1%), the probability that all rows will be removed by listwise deletion can remain extremely low even when $k$ is qualitatively large (e.g., $k = 150$) and $n$ is qualitatively small (e.g., $n = 100$). However, our results are striking once idiosyncratic missingness rates approach 10% or 25%, with striking consequences to the amount of data remaining following listwise deletion.

## 3 Application

In order to understand the real-world operating characteristics of listwise deletion, we considered two prominent datasets in use in the fields of comparative politics and international relations: the January 2021 QoG (Teorell *et al.* 2021) standard cross-sectional dataset, and the State Failure dataset covering country-years from 1955 to 1998 (King and Zeng 2007) reported by Esty *et al.* (1995, 1999) and considered by King and Zeng (2001). Table 1 provides summary statistics on these datasets. We applied mild preprocessing to these datasets: we removed country code variables,

[1] Lehmann (1999, 255) provides a good discussion in the context of a sample of size $n$ from a population of size $N$: "…we must go back to the purpose of embedding a given situation in a fictitious sequence: to obtain a simple and accurate approximation. The embedding sequence is thus an artifice and has only this purpose which is concerned with a particular pair of values $N$ and $n$ and which need not correspond to what we would do in practice as these values change." Our discussion is analogous, with our "pair of values" being $k$ and $n$. Thanks to Fredrik Sävje for suggesting the reference.

https://doi.org/10.1017/pan.2022.5 Published online by Cambridge University Press

Table 1. Summary statistics for QoG and State Failure datasets.

| | Quality of Government | State Failure |
|---|---|---|
| Units of observation | Countries | Country-years |
| Number of observations | 194 | 8, 580 |
| Number of variables | 351 | 1, 205 |
| Proportion of missing values (avg.) | 35.9% | 66.8% |
| Proportion of missing values (max) | 90.7% | 99.99% |
| Number of variables fully observed | 6 | 79 |

and in the case of the State Failure data, to apply the principle of charity, we removed 19 variables that exhibited 100% missingness.

## 3.1 Methodology

We conducted simulations that ask: how much data are lost by listwise deletion if we randomly subsample $k$ of the variables included in each of these datasets? Our simulation thus attempts to understand statistical behavior when using variables typical (or at least representative) of the major datasets in use in comparative politics and international relations. To do so, we conducted 25, 000 simulations in each of which we drew $k$ of the variables from each dataset (without replacement). We then recorded the number of rows of the data that survive listwise deletion.

We then report the expected proportion of remaining data observed after listwise deletion, as well as the probability that all rows of the data exhibit some missingness. Note that, here, expected values and probabilities refer to the randomness induced by our random subsampling procedure, not any fundamental stochasticity giving rise to the underlying data. Insofar as random sampling of variables from these datasets codifies a notion of representativeness, interpretation naturally follows from that notion.

We briefly discuss how out theoretical assumptions align with this setting. Assumption 1 asserts i.i.d. missingness across rows, which is unlikely to be met in this setting. In the cross-sectional QoG case, some countries (e.g., members of the European Union) have highly correlated missingness patterns, in part because of shared data availability. In the State Failure dataset, this is more dramatic: since observations are at the country-year, missingness is typically positively correlated within countries. The consequence—as in other clustering problems—is that the effective $n$ may be smaller than the nominal $n$ in practice. Our theoretical assumption of i.i.d. missingness across rows can therefore be seen as optimistic when faced with real-world data and, all else equal, the probability that all data will be lost may be higher than theory might dictate.

Assumption 2 asserts idiosyncratic missingness; that is, that there exists a factorization of the data such that all observations have some probability of missingness. Assumption 2 holds in our setting, because each observation in each dataset has missingness on at least one variable. Since our subsamples are formed via random sampling, if the first $k - 1$ randomly selected variables are not missing, it follows that the $k$th variable has a nonzero probability of being missing. Thus, our theoretical assumption of idiosyncratic missingness is compatible with the data, since it is met under a model of random sampling of variables.

## 3.2 Results

The left panel of Figure 1 plots the expected proportion of observed data against the number of randomly selected variables in the simulation for the QoG dataset. This proportion monotonically decreases to 0 as the number of randomly selected variables increases. With only 2 randomly

https://doi.org/10.1017/pan.2022.5 Published online by Cambridge University Press

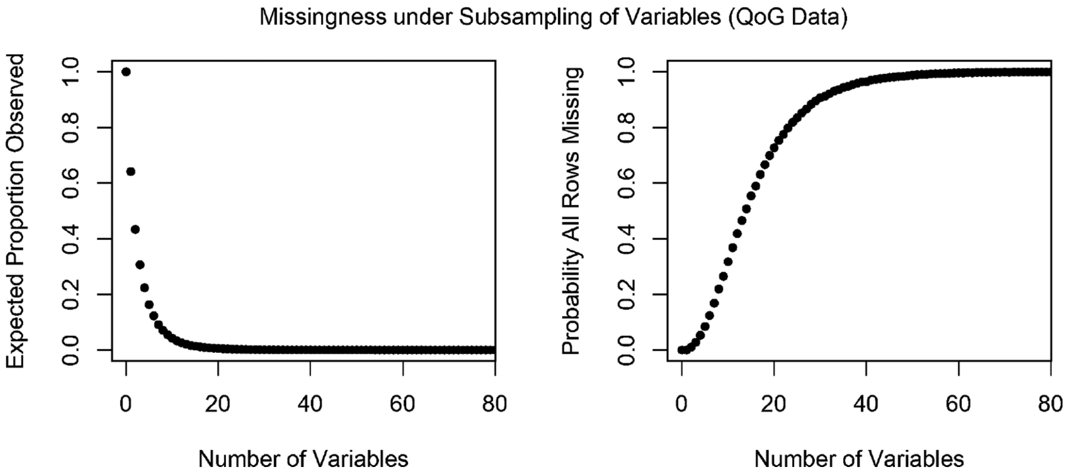

**Figure 1.** Properties of listwise deletion on random subsamples of variables from the QoG dataset.

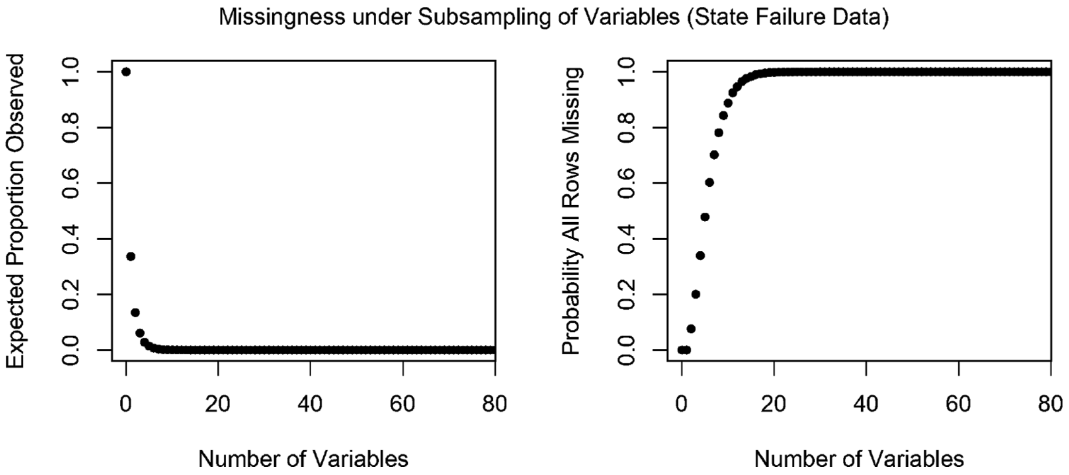

**Figure 2.** Properties of listwise deletion on random subsamples of variables from the State Failure dataset.

selected variables used, the researcher can expect to lose more than 50% of the data, which becomes more than 99% when 17 randomly selected variables are used.

The right panel of Figure 1 plots the probability of all rows will be unusable following listwise deletion against the number of randomly selected variables on the *x*-axis. Our theory would predict that this probability should converge to 1 as we increase the number of variables selected, and this is borne out in our simulation. The researcher can expect to lose all of the data with probability greater than 0.5 with 14 variables selected. With 52 variables used, this probability becomes more than 0.99.

We found similar, but more dramatic, trends with the State Failure dataset. The left panel of Figure 2 plots the expected proportion of observed data against the number of randomly selected variables for the State Failure dataset. The proportion decreases at a faster rate compared with the QoG dataset. The researcher can expect to lose more than 50% of the data if the only variable included is randomly selected. With more than three variables, the loss is more than 99%.

The right panel of Figure 2 plots the probability of all rows missing under listwise deletion against the number of randomly selected variables on the *x*-axis. This probability converges to 1 in a faster rate, as the researcher can expect to lose all of the data with probability greater than 0.5 with 6 variables included and the probability rises to be greater than 0.99 when including more than 17 variables. Taken together with our results from the QoG data, these results demonstrate that the moral of our theoretical results can be seen in real-world settings.

https://doi.org/10.1017/pan.2022.5 Published online by Cambridge University Press

## 4 Discussion

Our results demonstrate that listwise deletion cannot generally accommodate many variables, and that this problem is not resolved asymptotically. Application of high-dimensional asymptotics reveals that listwise deletion is even more fragile than was previously understood. Examining real-world data used in the fields of comparative politics and international relations highlights the seriousness of these issues for the types of data that political scientists use.

Our results imply that scholars who are committed to listwise deletion may be unable to use all of the variables that are necessary for an otherwise valid data analysis even when $n$ is large. For example, in order to achieve valid inferences in an observational study, a scholar may identify a large number of variables necessary to be conditioned on. However, if these variables exhibit idiosyncratic missingness, then the use of listwise deletion would require the scholar to exclude variables that would be necessary to attain an unbiased estimate. Neither dropping necessary variables nor dropping many observations is desirable. Approaches that avoid listwise deletion exist, including in the high-dimensional setting (e.g., Liu *et al.* 2016), and the researcher should consider these alternatives.

We conclude by emphasizing that this note should not be read as advocacy for the generic use of any particular method for addressing missing data. As Arel-Bundock and Pelc (2018) and Pepinsky (2018) demonstrate, no best method is best across all settings, and listwise deletion can outperform alternatives (e.g., multiple imputation) depending on the underlying data generating process. Our results provide additional support for the perspective that the most suitable inferential strategy is one chosen based on the specifics of the problem at hand.

## Appendix: Proofs

*Proof of Lemma 3.* We will prove the result in two cases. First suppose $q_* = 0$, which is equivalent to say that all variables are fully missing. Then $p_{all} = 1 = (1 - q_*^k)^n$. Now suppose $q_* \in (0, 1)$. By the independence assumption, $p_{all} = \Pi_{i=1}^{n}(1 - \Pr(\mathbf{M}_{ik} = 0))$. Denote $q_{ij} = \Pr(M_{ij} = 0 | \mathbf{M}_{i(j-1)} = 0)$ if $j > 1$, else $q_{ij} = \Pr(M_{ij} = 0)$. By Assumption 2, $q_{ij} \leq q_*$, for all $j \in \{1, 2, \ldots, k\}$. By the chain rule of conditional probability, $\Pr(\mathbf{M}_{ik} = 0) = q_{i1}, q_{i2}, \ldots, q_{ik}$. This means that the probability of a single observation containing at least one missing entry is $(1 - q_{i1}, q_{i2}, \ldots, q_{ik})$. Since $q_* \geq q_{ij}$ for all $j \in \{1, 2, \ldots, k\}$, $q_*^k \geq q_{i1}, q_{i2}, \ldots, q_{ik}$. Thus $(1 - q_*^k) \leq (1 - q_{i1}, q_{i2}, \ldots, q_{ik})$. Thus $(1 - q_*^k)^n$ is a lower bound for the probability of all $n$ observations each containing at least one missing entry. □

*Proof of Proposition 5.* First we will show that $\lim_{n\to\infty} nq_*^{f(n)} = 0$ (in asymptotic shorthand notation, $q_*^{f(n)} = o(n^{-1})$). Note that

$$\lim_{n\to\infty} nq_*^{f(n)} = \lim_{n\to\infty} e^{\log nq_*^{f(n)}} = \lim_{n\to\infty} e^{\log n + f(n)\log q_*}.$$

Since $q_* \in (0, 1)$, $\log q_* < 0$. Since $f(n) = \omega(\log n)$, the sequence $\log n + f(n) \log q_*$ diverges to negative infinity, and so

$$\lim_{n\to\infty} e^{\log n + f(n)\log q_*} = 0 = \lim_{n\to\infty} nq_*^{f(n)}.$$

Since $q_* \in (0, 1)$ and $k = f(n) \geq_* 1, -q_*^{f(n)} > -1$ and $1 - q_*^{f(n)} \leq 1$. By Bernoulli's Inequality, since $n \in \mathbb{N}, (1 - q_*^{f(n)})^n \geq_* 1 + n(-q_*^{f(n)}) = 1 - nq_*^{f(n)}$. Thus $1 - nq_*^{f(n)} \leq (1 - q_*^{f(n})^n \leq 1$ in the common domain $n \in \mathbb{N}$. Since $\lim_{n\to\infty} 1 = 1$ and $\lim_{n\to\infty} 1 - nq_*^{f(n)} = 1 - \lim_{n\to\infty} nq_*^{f(n)} = 1$, by the Squeeze Theorem,

$$\lim_{n\to\infty} (1 - q_*^{f(n)})^n = 1.$$

Then, since $\forall n, (1 - q_*^{f(n)})^n \leq p_{all} \leq 1$, we have $\lim_{n\to\infty} p_{all} = 1$, again by the Squeeze Theorem. □

https://doi.org/10.1017/pan.2022.5 Published online by Cambridge University Press

## Acknowledgment

The authors thank Forrest Crawford, Natalie Hernandez, Rosa Kleinman, Fredrik Sävje, as well as the Editor, Jeff Gill, and two anonymous referees for helpful comments.

## Data Availability Statement

Data and code to replicate all simulations and numerical illustrations are available at Wang and Aronow (2021).

## Supplementary Material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2022.5.

## References

Allison, P. D. 2001. *Missing Data*, Quantitative Applications in Social Sciences, Vol. 136. Thousand Oaks: Sage.

Arel-Bundock, V., and K. J. Pelc. 2018. “When Can Multiple Imputation Improve Regression Estimates?” *Political Analysis* 26 (2): 240–245.

Berk, R. 1983. “Applications of the General Linear Model to Survey Data.” In *Handbook of Survey Research*, edited by A. B. A. Peter, H. Rossi, and J. D. Wright, pp. 495–546. Quantitative Studies in Social Relations. New York: Academic Press.

Cameron, A., and P. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Esty, D. C., et al. 1999. “State Failure Task Force Report: Phase II Findings.” Environmental Change and Security Project Report 5: 49–72.

Esty, D. C., J. Goldstone, T. R. Gurr, P. Surko, and A. Unger. 1995. *Working Papers: State Failure Task Force Report*. McLean: Science Applications International Corporation.

Friedman, J., T. Hastie, and R. Tibshirani. 2010. “Regularization Paths for Generalized Linear Models via Coordinate Descent.” *Journal of Statistical Software* 33 (1): 1–22.

Honaker, J., and G. King. 2010. “What to Do About Missing Values in Time-Series Cross-Section Data.” *American Journal of Political Science* 54 (2): 561–581.

King, G., J. Honaker, A. Joseph, and K. Scheve. 2001. “Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation.” *American Political Science Review* 95: 49–69.

King, G., and L. Zeng. 2001. “Improving Forecasts of State Failure.” *World Politics* 53 (4): 623–658.

King, G., and L. Zeng. 2007. “Replication Data for: Improving Forecasts of State Failure.” Harvard Dataverse.

Lai, T. L., H. Robbins, and C. Z. Wei. 1978. “Strong Consistency of Least Squares Estimates in Multiple Regression.” *Proceedings of the National Academy of Sciences of the United States of America* 75 (7): 3034–3036.

Lall, R. 2016. “How Multiple Imputation Makes a Difference.” *Political Analysis* 24 (4): 414–433.

Lehmann, E. 1999. *Elements of Large-Sample Theory*, Springer Texts in Statistics. New York: Springer.

Little, R. J., and D. B. Rubin. 2019. *Statistical Analysis with Missing Data*, Wiley Series in Probability and Statistics, Vol. 793. Hoboken: Wiley.

Liu, Y., Y. Wang, Y. Feng, and M. M. Wall. 2016. “*Variable Selection and Prediction with Incomplete High-Dimensional Data*.” *The Annals of Applied Statistics* 10 (1): 418–450.

Pepinsky, T. B. 2018. “A Note on Listwise Deletion Versus Multiple Imputation.” *Political Analysis* 26 (4): 480–488.

R Core Team . 2020. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Schafer, J. L. 1997. *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC.

Stata.com . 2020. “Regress—Linear Regression.” https://www.stata.com/manuals13/rregress.pdf.

Teorell, J., A. Sundström, S. Holmberg, B. Rothstein, N. A. Pachon, and C. M. Dalli, 2021. “The Quality of Government Standard Dataset, Version Jan21.” University of Gothenburg, The Quality of Government Institute.

Wang, J. S., and P. M. Aronow. 2021. “Replication Data for: Listwise Deletion in High Dimensions.” Harvard Dataverse, Draft Version, UNF:6:0gB5c9RyKb6AH1zMEUNOpQ==[fileUNF].” https://doi.org/10.7910/DVN/T8BG2K.

https://doi.org/10.1017/pan.2022.5 Published online by Cambridge University Press