INFERENCE WITH IMPUTED DATA: The Allure of Making Stuff Up

Charles F. Manski
Department of Economics and Institute for Policy Research
Northwestern University

First public draft, May 2022. This version, June 2023.

Forthcoming, Journal of Labor Economics

Abstract

Incomplete observability of data generates an identification problem. What one can learn about a population parameter depends on the assumptions one finds credible. Rubin has promoted random multiple imputation (RMI) as a general way to deal with missing values. The recommendation has been influential to researchers who seek a simple fix to the nuisance of missing data. This paper provides a transparent assessment of the mix of Bayesian and frequentist thinking used by Rubin to argue for RMI. It evaluates random imputation to replace missing outcome or covariate data when the objective is to learn a conditional expectation.

Prepared for the Conference in Honor of Robert Moffitt, Johns Hopkins University, September 9-10, 2022.

1. Introduction

A classic concern of statistics is to use sample data to infer features of a conditional probability distribution. Consider a population characterized by a joint distribution P(y, x), where y is a real outcome and x is a covariate vector. The objective is to learn about P(y|x). One observes (y_i, x_i, i = 1, . . , N) in a random sample of N persons drawn from a study population that has distribution P(y, x). One uses the sample data to estimate features of P(y|x), often the conditional mean E(y|x). For example, a labor economist may want to learn the distribution of income or wages in households with specified covariates.

It is well-understood that incomplete observability of sample data generates an identification problem. Inference without assumptions requires contemplation of all logically possible distributions of the missing data. Doing so yields the set of all possible values of P(y|x), its identification region. The practical challenge is to characterize this set in a tractable way. Manski (1989, 1994) showed that identification analysis for E(y|x) and conditional quantiles is simple when only outcome data are missing. Analysis is more complex when the objective is to learn a spread parameter such as V(y|x); see Blundell et al. (2007) and Stoye (2010). Analysis is also more complex when some sample members have missing covariate data. Horowitz and Manski (1998, 2000) study these settings, with focus on E(y|x).

It is well-understood that assumptions about the distribution of missing data have identifying power. Relatively weak assumptions may shrink the identification region for P(y|x). Sufficiently strong assumptions may yield point identification. Manski (2003) assembles findings on identification using a spectrum of assumptions. A basic lesson is that there is no panacea for missing data. What one can learn about a population parameter depends on the assumptions one finds credible to maintain. The credibility of assumptions varies with the empirical setting under study. No specific assumptions can provide a realistic general solution to the problem of inference with missing data.

A common approach to coping with missing data is to impute them: the word “imputation” means using artificially constructed values, sometimes called “synthetic data,” to take the place of missing data.

Whether explicitly or implicitly, every imputation method uses assumptions about the distribution of missing data to generate the constructed values. The results depend critically on the assumptions made.

Illustration: The U.S. Census Bureau has applied *hot-deck* imputation to the Current Population Survey (CPS), describing the method this way (U. S. Census Bureau, 2006, p. 9-2):

> "This method assigns a missing value from a record with similar characteristics, which is the hot deck. Hot decks are defined by variables such as age, race, and sex. Other characteristics used in hot decks vary depending on the nature of the unanswered question. For instance, most labor force questions use age, race, sex, and occasionally another correlated labor force item such as full- or part-time status."

Thus, agency staff select a vector of covariates for which response is complete and compute the empirical distribution of the outcome of interest among sample members who have this covariate value and who report their outcomes. A hot-deck outcome is imputed to a person with missing data by drawing a realization at random from the available empirical distribution.

The CPS documentation of hot-deck imputation offers no evidence that the method yields an outcome distribution for missing data that is close to the actual distribution of such outcomes. The method used to generate imputations matters greatly when missing data is common. Consider, for example, the use of CPS data on income to estimate the U.S. family poverty rate. Table 1 of Manski (2016) shows that item nonresponse to income questions is substantial. Table 1 here is an excerpt from that table, presenting interval estimates of the poverty rate that make no assumption about the distribution of missing income data and point estimates using Census Bureau hot-deck imputations.

Table 1: Family Poverty Rate: Excerpt from Table 1 of Manski (2016)

Year	Interviewed Families	Fraction with Missing Data	Interval Estimate	Point Estimate with Imputations
2001	89063	0.436	[0.110, 0.315]	0.146
2011	86038	0.384	[0.139, 0.339]	0.176

The point estimates necessarily lie within the intervals because the imputations are logically possible values of the missing data. The point estimates could potentially lie anywhere within the intervals, depending on the imputation method used. □

While no single imputation method has been used universally, random multiple imputation (RMI) has been promoted as a general way to deal with missing values in public-use data (Rubin, 1987, 1996). The adjective "random" refers to drawing imputed values from a specified probability distribution. The adjective "multiple" refers to repetition of the random imputation process, generating multiple pseudo datasets and correspondingly multiple estimates of parameters of interest.

Rubin (1996) made this broad recommendation (p. 473): "For the context for which it was envisioned, with database constructors and ultimate users as distinct entities, I firmly believe that multiple imputation is the method of choice for addressing problems due to missing values." This recommendation has been influential. Considering missing data in clinical trials, a National Research Council panel (National Research Council, 2010) argued favorably for RMI. Use of RMI has also been recommended for observational studies in medicine (e.g., Sterne *et al.*, 2009; Azur *et al.,* 2011; Pedersen *et al.*, 2017).

Enthusiasm for RMI has extended beyond application to missing data to encompass its use to replace observed data that are deemed sensitive, the aim being to preserve the privacy of sample members. Rubin (1993) proposed that a data steward use observed data to estimate a model approximating the probability distribution of sensitive data conditional on non-sensitive data. One then uses random draws from this modelled distribution to replace the sensitive data. Repeating this process yields multiple imputations. The idea has subsequently been discussed by Reiter (2002), Drechsler and Reiter (2009), Hotz *et al.* (2022), and others. In research on data privacy, imputations are usually called synthetic data.

It is easy to see why imputation may be attractive to empirical researchers who seek a simple fix to the nuisance of missing data. One imputes data and then performs statistical analysis as if they are actual data. Rubin (1996) put it this way (p. 474): "Each tool in the ultimate users' existing arsenals can be applied to

any data set with missing values using the same command structure and output standards as if there were no missing data." This is the allure of making stuff up.

The allure is superficial. RMI is an operational procedure to approximate a modelled distribution of missing data by repeated simulation and to facilitate use of readily available statistical software. RMI is well-grounded only if the modelled distribution closely approximates the actual distribution of missing data, conditional on the observed data. However, there is rarely good reason to think that modelled distributions of missing data are accurate. It has been common to assume that data are missing at random and to use observed data to model the distribution of missing data. Analysts typically invoke this assumption without much if any comment about its credibility. A frank exception is a U.S. Census Bureau document describing the American Housing Survey, which states (U. S. Census Bureau, 2011):

> "Some people refuse the interview or do not know the answers. When the entire interview is missing, other similar interviews represent the missing ones . . . . For most missing answers, an answer from a similar household is copied. The Census Bureau does not know how close the imputed values are to the actual values."

Commentators have expressed concern about the accuracy of modelled distributions for sensitive data. Matthews and Harel (2011) observe that inferences made with imputed data are generically incorrect if the imputation model is incorrect. Reiter (2002) states (p. 532): "the validity of inferences depends critically on the accuracy of the imputation model." He explains that "When these models fail to reflect accurately certain relationships, analysts' inferences also will not reflect those relationships. Similarly, incorrect distributional assumptions built into the models will be passed on to the users' analyses." Hotz *et al.* (2022) write: "in cases where data sets have a large number of variables and the number of relationships that researchers might want to investigate is very large, it is essentially impossible for all relationships to be accurately captured by the type of models that are feasible to use for synthetic data set creation."

I find it disturbing that empirical researchers continue to trust in imputation in general, and RMI in particular, even though maintained assumptions about the distribution of missing data commonly lack foundation. The simplicity of imputation to the user, with its superficial enablement of conventional inference, does not suffice to justify its use in empirical research. Credibility should matter.

I have previously cautioned against poorly motivated imputation, notably in Horowitz and Manski (1998) and Manski (2016). Credible assumptions about the distribution of missing data are commonly too weak to yield point identification. Given this, I have recommended that empirical researchers acknowledge the realism of partial identification and report appropriate interval estimates. When point estimation is necessary as an input to decision making, I have recommended that statistical decision theory be used to derive estimates with desired properties (Dominitz and Manski, 2017).

This paper adds to my earlier work in two main ways. Section 2 provides a transparent assessment of the mix of Bayesian and frequentist thinking that has been used to argue for RMI. While it is not new research in the traditional sense, this assessment is an original contribution in the sense of providing an easily understandable new explanation of a subject whose discussion has often been opaque. In particular, I explain that the fundamental part of RMI is the distribution used to generate random imputations, not the generation of multiple imputations. The "multiple" aspect of RMI is simply a renaming of Monte Carlo integration to approximate the mean of a Bayesian posterior distribution for a parameter of interest.

Section 3, building on and adding to Horowitz and Manski (1998), evaluates random imputation to replace missing outcome or covariate data when the objective is to learn a conditional expectation. The primary new contribution is my study of the probability limits of imputations of outcomes (Section 3.1.1) and covariates (Section 3.2.1). This analysis reveals the specific distributional assumptions that must be satisfied if an estimator that imputes missing data is to converge asymptotically to a population parameter of interest. The concluding Section 4 considers steps that might help combat the allure of making stuff up.

This paper does not discuss the large literature that assumes a known distribution of missing data and then studies the finite-sample statistical imprecision of alternative imputation or other estimates. Examples include inverse probability weighting for missing outcome and covariate data (Wooldridge, 2007) and use of overidentifying restrictions for missing covariates (Abrevaya and Donald, 2017). This work is valuable when it is credible to assume a particular distribution of missing data, but my concern is the common situation in which a credible assumption is lacking. In this situation, the primary issues for analysis are identification of parameters and the consistency of estimates rather than their statistical precision.

2. The Bayesian Theory and Frequentist Interpretation of RMI

2.1. Bayesian Theory

RMI was originally motivated from a subjective Bayesian perspective. One places a joint subjective distribution on all observed and missing data. One computes the posterior subjective distribution of missing data conditional on the observed data. One uses this to derive the posterior distribution of a parameter of interest. Given this, RMI is simply a computational method that uses Monte Carlo integration to approximate the mean of the posterior distribution.

The Bayesian theory was developed in a series of articles and a book (Rubin, 1987). A concise statement was given in Rubin (1996), where he considered the posterior distribution for a real parameter Q(Y), Y being a random vector with some components observed and some missing. He wrote (p. 476):

> "The key Bayesian motivation for multiple imputation is given by result 3.1 in Rubin (1987). . . . . the results and its consequences can be easily stated using the simplified notation that the complete-data are $Y = (Y_{obs}, Y_{mis})$, where Y_{obs} is observed and Y_{mis} is missing. Specifically, the basic result is
>
> $$P(Q|Y_{obs}) = \int P(Q|Y_{obs}, Y_{mis})P(Y_{mis}|Y_{obs})dY_{mis}.$$"

Here $P(Q|Y_{obs})$ is the posterior predictive distribution of Q conditional on Y_{obs}, $P(Q|Y_{obs}, Y_{mis})$ is the posterior for Q given (Y_{obs}, Y_{mis}), and $P(Y_{mis}|Y_{obs})$ is the posterior for Y_{mis} given Y_{obs}. $P(Q|Y_{obs}, Y_{mis})$ and $P(Y_{mis}|Y_{obs})$ are specified subjective distributions, making $P(Q|Y_{obs})$ computable. In practice, the focus has been on the posterior mean of Q: $E(Q|Y_{obs}) = \int E(Q|Y_{obs}, Y_{mis})P(Y_{mis}|Y_{obs})dY_{mis}$.

What Rubin called "the basic result" does not refer to RMI. He interpreted it as RMI by considering Monte Carlo integration as a practical approach to approximate $E(Q|Y_{obs})$. One draws repeated values of Y_{mis} at random from $P(Y_{mis}|Y_{obs})$ and averages the resulting values of $E(Q|Y_{obs}, Y_{mis})$. Semantically, one may refer to Monte Carlo draws of Y_{mis} as imputations. Hence, RMI is Monte Carlo integration.

2.2. Frequentist Interpretation

The above motivation for RMI is well-grounded from a subjective Bayesian perspective. A disconnect between the theory and practice of RMI stems from efforts made to assert desirable frequentist properties for RMI. To a subjective Bayesian, the posterior mean $E(Q|Y_{obs})$ is interpretable regardless of whether it equals an objective quantity of scientific interest. A frequentist, however, assumes the existence of an objective quantity of interest, say Q^*, and wants to estimate this quantity well across repeated samples.

The posterior mean $E(Q|Y_{obs})$ need not be a good estimate of Q^* when $P(Q|Y_{obs}, Y_{mis})$ and $P(Y_{mis}|Y_{obs})$ are simply subjective distributions. To prove good frequentist properties for $E(Q|Y_{obs})$, one typically assumes that the distributions $P(Q|Y_{obs}, Y_{mis})$ and $P(Y_{mis}|Y_{obs})$ are objectively correct. Rubin (1996) demonstrated awareness of this core requirement when he wrote (p. 474): "My conclusion is that 'correctly' modeling the missing data must be, in general, the data constructor's responsibility." However, he provided no evidence that data constructors are able to model missing data correctly.

He argued that a desirable frequentist property for statistical procedures is "randomization validity," interpreted as requiring approximately unbiased point estimates of scientific estimands. He wrote (p. 476): "Multiple imputation was designed to satisfy both achievable objectives by using the Bayesian and frequentist paradigms in complementary ways: the Bayesian model based approach to *create* procedures, and the frequentist (randomization-based approach) to *evaluate* procedures." Continuing, he wrote that if the multiple imputations are "proper" and complete data inference is randomization-valid, then (p. 477): "the large-*m* repeated-imputation inference . . . is randomization-valid for the scientific estimand *Q, no matter how complex the survey design.*"

I find it difficult to understand the extended verbal discussion of "proper" multiple imputation. However, I believe that I understand the type of frequentist inference that he had in mind. His symbol *m* refers to the number of random draws made from $P(Y_{mis}|Y_{obs})$. Hence, "large-*m*" refers to asymptotic analysis as *m* goes to infinity. Thus, it means that Monte Carlo integration yields a well-behaved estimate of a population mean as the number of pseudo-draws goes to infinity. Randomization validity in this sense

means that RMI yields a consistent estimate of the subjective posterior mean $E(Q|Y_{obs})$, asymptotically in m. It implies nothing about the performance of RMI in estimation of the objective quantity Q^*.

3. Imputation Estimation of Conditional Expectations

To obtain a concrete sense of the behavior of imputation estimates, this section studies estimation of a conditional expectation when some data are missing and are replaced by imputations. Section 3.1 considers the simple case of imputation of missing outcomes and Section 3.2 the more subtle one of imputation of missing covariates. I focus throughout first on identification and then on the probability limit of estimates that use imputations. I make only a few remarks on statistical inference with finite-sample data.

3.1. Imputation of Missing Outcomes

Consider a population with members characterized by variables (y, x, z). Here y is a real outcome with closed domain Y and x is a covariate vector with finite domain X. Realizations of x are always observable, but some realizations of y are not. The binary variable z indicates whether y is observable (z = 1) or not (z = 0). The distribution of (y, x, z) is denoted P. The objective is to learn $E(y|x = \xi)$ when $P(x = \xi) > 0$.

A simple argument presented in Manski (1989) yields the identification region for $E(y|x = \xi)$. Use the Law of Iterated Expectations to write

$$(1) \quad E(y|x = \xi) = E(y|x = \xi, z = 1)P(z = 1|x = \xi) + E(y|x = \xi, z = 0)P(z = 0|x = \xi).$$

$E(y|x = \xi, z = 1)$ and $P(z = 1|x = \xi)$ are observable but $E(y|x = \xi, z = 0)$ is not. Knowledge of the domain Y restricts $E(y|x = \xi, z = 0)$ to lie in the interval $[Y_L, Y_U]$, where $Y_L \equiv \min(Y)$ and $Y_U \equiv \max(Y)$. Hence, the identification region with no assumptions on the distribution of missing data is the interval

(2) $[E(y|x = \xi, z = 1)P(z = 1|x = \xi) + Y_L P(z = 0|x = \xi),\ E(y|x = \xi, z = 1)P(z = 1|x = \xi) + Y_U P(z = 0|x = \xi)]$.

If assumptions restrict $E(y|x = \xi, z = 0)$ to a proper subset of $[Y_L, Y_U]$, say Γ, the identification region is

(3) $E(y|x = \xi, z = 1)P(z = 1|x = \xi) + \gamma \cdot P(z = 0|x = \xi), \gamma \in \Gamma$.

Example: To illustrate (2), Manski (1989) considered a missing-data problem that arose in a study of exit from homelessness undertaken by Piliavin and Sosin (1988). These researchers wished to learn the probability that an individual who was homeless at a given date would have a home six months later. Thus, the population of interest was the set of people who were homeless at the initial date. The variable y was binary, with y = 1 if the individual had a home six months later and y = **0** if he or she remained homeless. The regressors x were individual background attributes. The objective was to learn $E(y|x) = P(y = 1|x)$. The investigators interviewed a random sample of the people who were homeless in Minneapolis in late December 1985. Six months later they attempted to reinterview the original respondents but succeeded in locating only a subset. Thus, the missing data problem was attrition from the sample: $z = 1$ if a respondent was located for reinterview, $z = 0$ otherwise. □

Random imputation estimates assume that $P(y|x = \xi, z = 0)$ is a specified distribution and use realizations drawn from this distribution to replace missing values of y. Suppose that a random sample of N population members are drawn. One observes (x_i, z_i) for all $i = 1, \ldots, N$ and observes y_i when $z_i = 1$. If y were always observed, one might naturally estimate $E(y|x = \xi)$ by the sample average $E_N(y|x = \xi,)$. To cope with missing outcome data, consider replacement of missing values of y with imputations and computation of the sample average combining observed and imputed data.

I now examine the probability limit of the imputation estimate as sample size goes to infinity, showing how the limit depends on the probability distribution used to generate imputations. It suffices to study the

limiting behavior of estimation using one synthetic sample. Multiple imputation yields multiple synthetic samples and multiple estimates, each with the same probability limit.

3.1.1. Analysis

Let each member of the population be assigned an imputed outcome $u \in Y$, which is used to replace missing outcome data. In a sample of size N, Let $N(1, \xi)$ be the sub-sample where $(z = 1, x = \xi)$ and let $N(0, \xi)$ be the sub-sample where $(z = 0, x = \xi)$. Let $N_{1\xi} = |N(1, \xi)|$, $N_{0\xi} = |N(0, \xi)|$, and $\pi_{N\xi} \equiv N_{1\xi}/(N_{1\xi} + N_{0\xi})$. Whenever $N_{1\xi} > 0$ and $N_{0\xi} > 0$, the imputation estimate of $E(y|x = \xi)$ is

$$(4)\quad \theta_{N\xi} \equiv \frac{1}{N_{1\xi} + N_{0\xi}} \left(\sum_{i \in N(1, \xi)} y_i + \sum_{i \in N(0, \xi)} u_i \right)$$

$$= \pi_{N\xi} \frac{1}{N_{1\xi}} \sum_{i \in N(1, \xi)} y_i + (1 - \pi_{N\xi}) \frac{1}{N_{0\xi}} \sum_{i \in N(0, \xi)} u_i .$$

Let $N \to \infty$. The probability limit of $\theta_{N\xi}$ is

$$(5)\quad \theta_{\xi} \equiv E(y|x = \xi, z = 1) \cdot P(z = 1|x = \xi) + E(u|x = \xi, z = 0) \cdot P(z = 0|x = \xi).$$

In general, $\theta_{\xi} \neq E(y|\, x = \xi)$. Comparison of (5) with (1) shows that the imputation estimate is consistent if and only if

$$(6)\quad E(u|x = \xi, z = 0) = E(y|x = \xi, z = 0).$$

Seeking to justify (6), researchers sometimes assume that data are missing at random conditional on x and aim to draw imputations from a consistent estimate of the observable distribution $P(y|x, z = 1)$, say $P_N(y|x, z = 1)$. Thus, let

(7a) $P(y|x = \xi, z = 0) = P(y|x = \xi, z = 1)$,

(7b) $P(u|x = \xi, z = 0) = P_N(y|x = \xi, z = 1)$.

Equation (7a) is an untestable assumption. Given (7a), equation (7b) asymptotically implies (6).

3.1.2. Alternatives to Random Imputation

Alternatives to random imputation provide superior approaches to inference on $E(y|x)$. First suppose that one thinks it credible to assume that missing data have a particular distribution, say $Q(y|x = \xi, z = 0)$. Let $E_Q(y|x = \xi, z = 0)$ be the mean outcome under Q. Then a simple alternative to random imputation is to replace missing values by $E_Q(y|x = \xi, z = 0)$, yielding the estimate

$$(8)\quad \theta_{QN\xi} \equiv \pi_{N\xi} \frac{1}{N_{1\xi}} \sum_{i \in N(1,\,\xi)} y_i + (1 - \pi_{N\xi}) \cdot E_Q(y|x = \xi, z = 0).$$

Estimate (8) has the same probability limit as the imputation estimate (4). Moreover, it has greater statistical precision. The difference between a random imputation estimate and (8) is that a random imputation estimate uses a realization drawn from distribution Q. In contrast, (8) directly uses the distribution itself.

Now suppose that one lacks a credible basis to specify a distribution for missing data and contemplates estimation without assumptions. Interval (2) gives the identification region for $E(y|x = \xi)$. This interval is bounded if Y_L and Y_U are finite. Then a natural interval estimate for $E(y|x = \xi)$ is the sample analog of (2), namely

$$(9)\quad \left[\pi_{N\xi}\frac{1}{N_{1\xi}}\sum_{i\in N(1,\xi)} y_i + (1-\pi_{N\xi})Y_L,\ \pi_{N\xi}\frac{1}{N_{1\xi}}\sum_{i\in N(1,\xi)} y_i + (1-\pi_{N\xi})Y_U\right],$$

whose probability limit is (2). An example is the estimate of a bound on exiting homelessness in Manski (1989), discussed above.

Should it be necessary to provide a point estimate of E(y|x), an attractive option is to use the midpoint of interval (9). This estimate converges to the midpoint of (2), which minimizes the maximum value of asymptotic squared bias among all point estimates of E(y|x). Dominitz and Manski (2017) study the finite-sample performance of the midpoint estimate from the perspective of statistical decision theory and derive the maximum value of mean square error across all distributions of missing data.

3.2. Imputation of Missing Covariates

In practice, many patterns of missing data may occur within a covariate vector. Analysis of every possible pattern requires cumbersome notation, so I focus on settings in which some covariates are always observed whereas others may have missing data. I denote the former covariates as x and the latter as w. Thus, consider a population with members characterized by variables (y, x, w, z). Here y is a real outcome with domain Y, whereas x and w are covariate vectors with finite domains X and W. Realizations of (y, x) are always observable, but some realizations of w are not. The binary variable z now indicates whether w is observable (z = 1) or not (z = 0). The population distribution of (y, x, w, z) is P. The objective is to learn $E(y|x = \xi, w = \omega)$ when $P(x = \xi, w = \omega) > 0$.

Horowitz and Manski (1998) derived the identification region for E(y|x, w) with no assumptions on the distribution of missing data. The derivation is more subtle than with missing outcome data and the general form of the region is more complex than (2). However, Horowitz and Manski (2000) and Manski

(2003) show that the region has a simple explicit form when y is a binary outcome, say taking the values 0 and 1. Applying Manski (2003, Corollary 3.8.1), the identification region for $E(y|x = \xi, w = \omega)$ is the interval

$$(10)\quad \left[\frac{P(y=1, x=\xi, z=1, w=\omega)}{P(x=\xi, z=1, w=\omega) + P(y=0, x=\xi, z=0)}, \frac{P(y=1, x=\xi, z=1, w=\omega) + P(y=1, x=\xi, z=0)}{P(x=\xi, z=1, w=\omega) + P(y=1, x=\xi, z=0)}\right].$$

The lower bound is achieved if the distribution of missing data has $P(w = \omega|y = 1, x = \xi, z = 0) = 0$ and $P(w = \omega|y = 0, x = \xi, z = 0) = 1$. The upper bound is achieved if the distribution of missing data has $P(w = \omega|y = 1, x = \xi, z = 0) = 1$ and $P(w = \omega|y = 0, x = \xi, z = 0) = 0$.

Example: Horowitz and Manski (2000) illustrated a more general form of bound (10), applicable when data are sometimes missing for outcomes and sometimes for covariates. We considered the setting of Manski *et al.* (1992), which used data from the 1979 National Longitudinal Study of Youth (NLSY) to study the rate of high school graduation of youth in intact and non-intact families. The outcome took the value y = 1 if a youth received a high school diploma by 1985 and y = 0 otherwise. The covariates included x = (sex, race/ethnicity, family structure during adolescence), and w = (mother's and father's years of schooling). The pattern of missing data was complex. Some had missing outcome data, some had missing schooling data for one or more parent, and some had missing data for both the outcome and one or more parent. □

Random imputation estimates assume that $P(w|y, x = \xi, z = 0)$ is a specified distribution and use realizations drawn from this distribution to replace missing values of w. Although Rubin supposed that a researcher knows this distribution, I find it difficult to see how a researcher would know it in practice. Labor economists sometimes find it credible to know properties of a distribution P(y|x, w), such as an income or wage distribution. The distribution $P(w|y, x = \xi, z = 0)$, however, is an unusual object that plays no clear role in economic analysis.

Suppose that a random sample of N population members are drawn. One observes (y_i, x_i, z_i) for all i = 1, . . . , N and observes w_i when $z_i = 1$. If w were always observed, one might naturally estimate $E(y|x = \xi,$

$w = \omega$) by the sample average $E_N(y|x = \xi, w = \omega)$. To cope with missing covariate data, consider replacement of missing values of w with imputations and computation of the sample average combining observed and imputed data.

I now examine the probability limit of the estimate as sample size goes to infinity, showing how the limit depends on the probability distribution used to generate imputations. It again suffices to study the limiting behavior of estimation using one synthetic sample as multiple imputation yields multiple estimates, each with the same probability limit.

3.2.1. Analysis

Let each member of the population be assigned an imputed value $u \in W$, which is used to replace missing covariate data. In a sample of size N, Let $N(1, \xi, \omega)$ be the sub-sample where $(z = 1, x = \xi, w = \omega)$ and let $N_m(0, \xi, \omega)$ be the sub-sample where $(z = 0, x = \xi, u = \omega)$. Let $N_{1\xi\omega} = |N(1, \xi, \omega)|$, $N_{0\xi\omega} = |N(0, \xi, \omega)|$, and $\pi_{N\xi\omega} \equiv N_{1\xi\omega}/(N_{1\xi\omega} + N_{0\xi\omega})$. Then, when $N_{1\xi\omega} + N_{0\xi\omega} > 0$, the imputation estimate of $E(y|x = \xi, w = \omega)$ is

$$(11)\quad \theta_{N\xi\omega} \equiv \frac{1}{N_{1\xi\omega} + N_{0\xi\omega}} \left(\sum_{i \in N(1, \xi, \omega)} y_i + \sum_{i \in N(0, \xi, \omega)} y_i \right)$$

$$= \pi_{N\xi\omega} \frac{1}{N_{1\xi\omega}} \sum_{i \in N(1, \xi, \omega)} y_i + (1 - \pi_{N\xi\omega}) \frac{1}{N_{0\xi\omega}} \sum_{i \in N(0, \xi, \omega)} y_i .$$

Let $N \rightarrow \infty$. The probability limit of $\theta_{N\xi\omega}$ is

$$(12)\quad \theta_{\xi\omega} \equiv E(y|x = \xi, w = \omega, z = 1)\cdot\pi_{\xi\omega} + E(y|x = \xi, u = \omega, z = 0)\cdot(1 - \pi_{\xi\omega}),$$

where

$$(13)\quad \pi_{\xi\omega} = \frac{P(z = 1, x = \xi, w = \omega)}{P(z = 1, x = \xi, w = \omega) + P(z = 0, x = \xi, u = \omega)}$$

$$= \frac{P(z = 1, w = \omega|x = \xi)}{P(z = 1, w = \omega|x = \xi) + P(z = 0, u = \omega|x = \xi)}.$$

In general, $\theta_{\xi\omega} \neq E(y|x = \xi, w = \omega)$. By the Law of Iterated Expectations,

$$(14)\quad E(y|x = \xi, w = \omega) = E(y|x = \xi, w = \omega, z = 1)\cdot P(z = 1|x = \xi, w = \omega) + E(y|x = \xi, w = \omega, z = 0)\cdot P(z = 0|x = \xi, w = \omega).$$

Comparison of (12) and (14) shows that they coincide if

$$(15a)\quad P(z = 0, u = \omega|x = \xi) = P(z = 0, w = \omega|x = \xi),$$

$$(15b)\quad E(y|x = \xi, u = \omega, z = 0) = E(y|x = \xi, w = \omega, z = 0).$$

These equalities generally do not hold.

Equations (15a)-(15b) do hold if the distribution of imputations is

$$(16)\quad P(u|y, x, z = 0) = P(w|y, x, z = 0).$$

Multiplying both sides of (16) by the observable distribution $P(y, x, z = 0)$ yields

$$(17)\quad P(y, x, u, z = 0) = P(y, x, w, z = 0),$$

which implies (15a)-(15b). The problem, of course, is that satisfaction of equation (16) requires knowledge of the distribution $P(w|y, x, z = 0)$ of missing data.

Seeking to justify (16), researchers sometimes assume that w data are missing at random conditional on (y, x) and aim to draw imputations from a consistent estimate of the observable distribution P(w|y, x, z = 1), say $P_N(w|y, x, z = 1)$. Thus, let

(18a) $P(w|y, x, z = 0) = P(w|y, x, z = 1)$,

(18b) $P(u|y, x, z = 0) = P_N(w|y, x, z = 1)$.

Equation (18a) is an untestable assumption. Given (18a), equation (18b) asymptotically implies (16).

3.2.2. Imputation as an Attempted Solution to the Ecological Inference Problem

A polar case of missing covariates that warrants special attention occurs when w is always missing; thus, P(z = 0) = 1. Then no conclusions about $E(y|x = \xi, w = \omega)$ can be drawn without further information. The literature on *ecological inference* has studied settings in which further information arises from a separate sampling process that yields observations of (w, x) but not of y. Here one faces the problem of identification of P(y|x, w) given observability of P(y|x) and P(w|x).

Duncan and Davis (1953) considered identification of $P(y = 1|x = \xi, w = \omega)$ when y is a binary outcome. They sketched a proof that the identification region is the interval

$$(19) \qquad [0, 1] \cap \left[\frac{P(y = 1|x = \xi) - P(w \neq \omega|x = \xi)}{P(w = \omega|x = \xi)}, \frac{P(y = 1|x = \xi)}{P(w = \omega|x = \xi)}\right].$$

Horowitz and Manski (1995) formalized this finding and studied identification of $E(y|x = \xi, w = \omega)$ when y is real-valued. The latter analysis is more subtle than when y is binary. The identification region is an interval that does not have an explicit form but can be computed numerically. Cross and Manski (2002) extend the analysis. Ridder and Moffitt (2007) and Cho and Manski (2008) review aspects of the literature.

Some medical researchers have used imputation of genotypes in an attempt to perform ecological inference. In this setting, y is a patient outcome while (x, w) are genetic markers. One dataset yields observations of (y, x) and another provides observations of (x, w), enabling estimation of P(y|x) and P(w|x). For each patient i in the former dataset, the estimate of distribution $P(w|x = x_i)$ is used to impute w_i, yielding (y_i, x_i, u_i), i = 1, . . . , N. This partially synthetic dataset is analyzed using the imputations as if they were real covariate data. See Gragert *et al.* (2014), Tinckam *et al.* (2016), Geneugelijk *et al.(*2017), Kamoun *et al.* (2017), and Nilsson *et al.* (2019).

Manski et al. (2019) and Manski et al. (2021) counsel against this use of genotype imputation. By construction, imputations drawn from P(w|x) are statistically independent of actual outcomes y. Applying (12), the probability limit of the imputation estimate of $E(y|x = \xi, w = \omega)$ is $E(y|x = \xi)$. Thus, imputation of w accomplishes nothing.

3.2.3. Alternatives to Random Imputation

Suppose one thinks it credible to assume that missing w data have a particular distribution, say $Q(w|y, x = \xi, z = 0)$. This assumption can be used to estimate E(y|x, w) without constructing randomly imputed data. Use the Law of Total Probability to write

$$(20) \quad P(y, x, w) = P(y, x, w|z = 1)P(z = 1) + P(y, x, w|z = 0)P(z = 0)$$
$$= P(y, x, w|z = 1)P(z = 1) + P(w|y, x, z = 0)P(y, x|z = 0P(z = 0).$$

Distributions P(y, x, w|z = 1), P(y, x|z = 0), and P(z) are observable. Each is consistently estimable by its sample analog, denoted $P_N(y, x, w|z = 1)$, $P_N(y, x|z = 0)$, and $P_N(z)$. Inserting these estimates into (20), and assuming that Q is the distribution of missing covariates, yields a consistent estimate of P(y, x, w), namely

$$(21) \quad P_N(y, x, w) = P_N(y, x, w|z = 1)P_N(z = 1) + Q(w|y, x, z = 0)P_N(y, x|z = 0P_N(z = 0).$$

The conditional mean of $P_N(y, x, w)$ is a consistent estimate of $E(y|x, w)$.

Suppose that one lacks a credible basis to specify a distribution for missing data and contemplates estimation without assumptions. The complex general form of the identification region for $E(y|x = \xi, w = \omega)$ makes estimation of this region complex as well. However, estimation is easy when y is a binary outcome. Then interval (10) gives the identification region. A natural interval estimate is its sample analog. An example is the estimate of the bound on the rate of high school graduation in Horowitz and Manski (2000), discussed above.

Should it be necessary to provide a point estimate, an attractive option is to use the midpoint of this interval estimate. As N increases, this converges to the midpoint of (10), which minimizes the maximum value of asymptotic squared bias among all point estimates.

4. Combatting the Allure of Making Stuff Up

The use of imputed data is a striking illustration of research with incredible certitude (Manski, 2011, 2020). Arguing for RMI, Rubin (1996) wrote (p. 473): "alternative methods either require special knowledge and techniques not available to typical users or produce answers that are generally not statistically valid for scientific estimands." Development of user-friendly methods is a worthy objective, provided that the methods yield useful findings.

RMI and other imputation methods are useful only if the assumed distribution of missing data is close to correct. Rubin recognized this central requirement in principle when he wrote (p. 474): "My conclusion is that 'correctly' modeling the missing data must be, in general, the data constructor's responsibility." Yet assigning responsibility to the data constructor is futile if demonstrably correct modeling is not achievable. Assumed distributions of missing data commonly lack credible foundations. Hence, assertions that RMI yields findings that are valid in certain Bayesian and frequentist senses should not comfort empirical researchers who want to make credible inferences about the real world. The new analysis in Sections 3.1.1

and 3.2.1 of this paper shows the specific distributional assumptions that must be satisfied if an estimator that imputes missing data is to converge asymptotically to a population parameter of interest.

I have focused the analysis of Section 3 on the problem of learning a conditional mean because this is so often the objective of empirical research and because this problem is relatively easy to study. Extension of the analysis to other settings would be welcome. Random imputation appears to have especially severe difficulties when data are longitudinal, as sequences of repeated imputations become necessary. Hotz et al. (2022) call attention to this matter in the setting of data privacy, writing (Appendix B4-B5):

> "Synthetic data may be produced for waves of a longitudinal data set, using the existing information through wave t to form the predictive synthetic data conditional on data collected through that wave. But, how should one synthesize data for future waves of the survey? To be consistent, should one condition the predictive distribution to wave t + 1 on the synthetic data or on the confidential data through wave t? Conditioning the predictive distribution on the former will perpetuate any model misspecification in earlier waves. Conditioning it on the confidential data from previous waves makes use of the cumulative information in that data, but risks producing inconsistency in individual-level time series across the waves of the synthetic data."

The identification problem generated by missing data in longitudinal studies has been analyzed in Horowitz and Manski (1998, 2000).

Recognition of the fragility of imputation is necessary to combat the allure of making stuff up, but I doubt that it will suffice. Another necessary step is to provide tractable methods that enable credible empirical research. Section 3.1 showed that assumption-free interval estimation is simple with missing outcome data. Section 3.2 showed that it is simple with missing covariate data when y is a binary outcome.

Interval estimation using various potentially credible assumptions with identifying power is straightforward. Examples include monotone-instrumental-variable and bounded-variation assumptions (Manski and Pepper, 2000; Manski et al. 2019). On the other hand, identification analysis in some problems with missing data is complex. Empirical researchers should be sophisticated enough to recognize that performing credible analysis may be a challenge.

References

Abrevya, Jason and Stephen Donald (2017). A GMM approach for dealing with missing data on regressors. *The Review of Economics and Statistics* 99: 657-662.

Azur, Melissa, Elizabeth Stuart, Constantine Frangakis, and Philip Leaf (2011). Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research* 20: 40-49.

Blundell, Richard, Amanda Gosling, Hidehiko Ichimura, and Costas Meghir (2007). Changes in the distribution of male and female wages accounting for employment composition using bounds. *Econometrica* 75: 323-363.

Cho, Wendy, and Charles Manski (2008). Cross level/ecological inference. in *Oxford Handbook of Political Methodology.* ed. H. Brady, D. Collier, and J. Box-Steffensmeier. Oxford: Oxford University Press.

Cross, Philip, and Charles Manski (2002). Regressions, short and long. *Econometrica* 70: 357-368.

Dominitz, Jeff, and Charles Manski (2017). More data or better data? a statistical decision problem. *Review of Economic Studies* 84: 1583-1605.

Drechsler, Jörg and Jerry Reiter (2009). Disclosure risk and data utility for partially synthetic data: an empirical study using the German IAB establishment survey. *Journal of Official Statistics* 25: 589-603.

Duncan Otto, and Beverly Davis (1953). An Alternative to Ecological Correlation. *American Sociological Review* 18: 665-666.

Gragert, Loren, Stephanie Fingerson, Mark Albrecht, Martin Maiers, Matt Kalaycio, and Brian Hill (2014). Fine-mapping of HLA associations with chronic lymphocytic leukemia in US populations. *Blood* 124: 2657-2665.

Geneugelijk, Kristen, Jeroen Wissing, Dirk Koppenaal, Matthias Niemann, and Eric Spierings (2017). Computational approaches to facilitate epitope-based hla matching in solid organ transplantation. *Journal of Immunology Research*, https://doi.org/10.1155/2017/9130879.

Horowitz, Joel, and Charles Manski (1995). Identification and robustness with contaminated and corrupted data. *Econometrica* 63: 281-302.

Horowitz, Joel, and Charles Manski (1998). Censoring of outcomes and regressors due to survey nonresponse: identification and estimation using weights and imputations. *Journal of Econometrics* 84: 37–58.

Horowitz, Joel, and Charles Manski (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association* 95: 77–84.

Hotz, V. Joseph, Christopher Bollinger, Tatiana Komarova, Charles Manski, Robert Moffitt, Denis Nekipelov, Bruce Spencer, and Aaron Sojourner (2022). Balancing data privacy and usability in the federal statistical system. *Proceedings of the National Academy of Sciences* 119, https://www.pnas.org/doi/full/10.1073/pnas.2104906119.

Kamoun, Maloun, Keith McCullough, Martin Maiers, Marcelo Fernandez Vina, Hongzhe Li, Valerie Teal, Alan Leichtman, and Robert Merion (2017). HLA amino acid polymorphisms and kidney allograft survival. *Transplantation* 101: e170–e177.

Manski, Charles (1989). Anatomy of the selection problem. *Journal of Human Resources* 24: 343–360.

Manski, Charles (1994). The selection problem, in *Advances in Econometrics, Sixth World Congress*. ed. C. Sims, Cambridge: Cambridge University Press.

Manski, Charles (2003). P*artial identification of probability distributions.* New York: Springer-Verlag.

Manski, Charles (2011). Policy analysis with incredible certitude. *The Economic Journal* 121: F261-F289.

Manski, Charles (2016). Credible interval estimates for official statistics with survey nonresponse. *Journal of Econometrics* 191: 293-301.

Manski, Charles (2020). The lure of incredible certitude. *Economics and Philosophy* 36: 216-245.

Manski, Charles, and John Pepper (2000). Monotone instrumental variables: with an application to the returns to schooling. *Econometrica* 68: 997-1010.

Manski, Charles, Gary Sandefur, Sara McLanahan, and Daniel Powers (1992). Alternative estimates of the effect of family structure during adolescence on high school graduation. *Journal of the American Statistical Association* 87: 25-37.

Manski, Charles, Anat Tambur, and Michael Gmeiner (2019). Predicting kidney transplant outcomes with partial knowledge of HLA mismatch. *Proceedings of the National Academy of Sciences* 116: 20339-20345.

Manski, Charles, Michael Gmeiner, and Anat Tambur (2021). misguided use of observed covariates to impute missing covariates in conditional prediction: a shrinkage problem. https://arxiv.org/abs/2102.11334

Matthews, Gregory, and Ofer Harel (2011). Data confidentiality: a review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys* 5: 1-29.

National Research Council (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington, DC: The National Academies Press.

Nilsson, Johan, David Ansari, Mattias Ohlsson, Peter Höglund, Anna-Sophie Liedberg, J. Gustav Smith, Pierre Nugues, and Bodil Andersson (2019), Human leukocyte antigen-based risk stratification in heart transplant recipients—implications for targeted surveillance. *Journal of the American Heart Association*, 8, https://doi.org/10.1161/JAHA.118.011124.

Piliavin, Irving, and Michael Sosin (1988). Exiting homelessness: some recent empirical findings. Madison: Institute for Research on Poverty, University of Wisconsin.

Reiter, Jerry (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics* 18: 531-543.

Ridder, Geert, and Robert Moffitt (2007). The econometrics of data combination. In *Handbook of Econometrics* Vol. 6B, ed. J. Heckman and E. Leamer, Amsterdam: Elsevier.

Rubin, Donald (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Rubin, Donald (1993). Statistical disclosure limitation. *Journal of Official Statistics* 9: 461-468.

Rubin, D. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association* 91: 473-489.

Stoye, Jörg (2010). Partial identification of spread parameters. *Quantitative Economics* 1: 323-357.

Tinckam, Kathryn, Caren Rose, Sandaram Hariharan, and John Gill (2016). Re-examining risk of repeated HLA mismatch in kidney transplantation. *Journal of the American Society of Nephrology* 27: 2833-2841.

U. S. Census Bureau (2006). *Current Population Survey Design and Methodology.* Technical Paper 66, Washington, DC: U. S. Census Bureau.

U.S. Census Bureau (2011). *Current Housing Reports, Series H150/09, American Housing Survey for the United States: 2009*, Washington, DC: U.S. Government Printing Office.

Wooldridge, Jeffrey (2007). Inverse Probability Weighted Estimation for General Missing Data Problems. *Journal of Econometrics* 141: 1281-1301.