There's No Such Thing as a Free Lunch: The Case of Hot Deck Imputations and Sampling Variance

Christopher R. Bollinger[1]

May 2020

[1]Christopher R. Bollinger, Sturgill Professor of Economics, University of Kentucky, Lexington, KY 40506. crboll@uky.edu.
This research was conducted while the author was on sabbatical at the Institute for Social and Economic Research at the University of Essex. I thank the Leverhulme Foundation for financial support. I thank Barry Hirsch, Tom Crossley, Peter Lynn, Charles Hokayem, Carlos Lamarche, Amitabh Chandra, Ken Troske, Jim Ziliak, Ana Herrara, John Pepper and Dan Black for valuable comments and suggestions.

Abstract

Prior work has shown that the use of Hot Deck type imputations may induce bias when the imputation is used as the dependent variable in a regression setting. Some researchers have argued that imputations are useful in order to increase sample size when data are missing. This paper demonstrates that when the imputations occur as the dependent variable, the sampling variance when imputations are included is larger than both the sampling variance of the complete case (no imputations estimtor) and the sampling variance typically estimated by computer packages. The use of imputations in the dependent variable do not improve precision.

1 Introduction

Missing data is a significant issue in a variety of empirical applications. A common approach to missing data is to impute a value, often based upon other non-missing data provided from the same observation. Perhaps the most common example of this are the "hot deck" imputations provided by the U.S. Census Bureau for earnings in the Outgoing Rotation Group (ORG) measure of hourly earnings and the Annual Social and Economic Survey (ASEC) measure of annual earnings. This paper examines the efficiency of using hot deck type imputations in both sample mean estimation and most importantly in a regression setting when the imputed variable is used as a dependent variable. Other Census products and many other publicly released data sets also include hot deck type imputation. Other imputations will have similar impacts on sampling distributions and must be worked out in a case-by-case basis.

Little and Rubin (2014) provide an extensive treatise on imputation. The emphasis in their book is on obtaining consistent estimates and they often focus on estimation when missing data would bias results. While they provide a variety of options and include approaches to address proper estimation of standard errors, in many cases the standard errors require extensive knowledge of how imputation was conducted. They seldom compare estimation using imputations to estimation using only the observed sample, and generally assume - when discussing sampling variance - that imputation will be constructed by the user. These assumption are important and Little and Rubin (2014) make it clear that any imputation must be conducted in the context of the model to be estimated. Little and Rubin (2014) cite Dempster and Rubin (1983) and warn:

> "The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be

legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases."

However, data sets are often released with imputations built in. Researchers often give little thought to imputation and use the imputed observations as though they are "real" data. The release of data with imputations unfortunately encourages such behavior. Hirsch and Schumacher (2004) and Bollinger and Hirsch (2006) present arguments for why use of imputations produced by hot deck type approaches may bias regression results. Many researchers - presumably concerned about sample size - use imputations. Most inference using imputations fails to account for the additional sampling variance induced by imputations.

This paper examines simple cases primarily focused on a regression setting where the missing data are only in the dependent, Y, variable. In that case, the assumption of missing at random (missing rates related only to X values in the regression) combined with typical model assumptions result in estimation that is consistent for both the imputed sample (the sample containing a mix of observed and imputed data) and the complete case (no imputations) sample. In cases where either the missing at random assumption fails, or the model assumptions fail, far more complicated approaches would need to be taken than are typically used.

The results indict the use of imputations: far from improving precision they increase sampling variance and result in less precise estimates. The typical standard errors estimated by computer packages are not correct when imputations are used. Unless the researchers take care to adjust standard errors appropriately, inference will be biased. Researchers often include imputations because the increase sample sizes. The results here show that for missing dependent variables in a regression setting this is illusory, that the gain is due to improper estimates of the actual standard errors, not a gain in sample size.

Section two defines terms and establishes assumptions. Section three examines the case of hot deck type imputations in two settings: a simple mean with missing

completely at random (MCAR) and a regression setting with missing at random (MAR). Section four presents three simulation exercises which highlight the analytic results. Section five presents a short empirical example using March 2018 ASEC data in a standard Mincer wage equation. The conclusions provides suggestions for alternative approaches including Inverse Probability Weighting (Moffitt, Fitzgerald, Gottschalk, 1999; Wooldridge, 2002; 2007).

2 Missingness Model

For simplicity, I focus on data with a single Y variable and a covariate X. Inference is focused on some mean of Y, either the mean $E[Y] = \mu$ or the conditional expectation $E[Y|X] = \alpha + \beta X$. This paper is focused on missingness in Y and is motivated by the high rate of earnings missingness in the March Current Population Survey (see Bollinger et al. 2019; Bollinger and Hirsch, 2006; and Hirsch and Schumacher, 2004). Missing rates for the earnings data in the Outgoing Rotation Groups are over 30%. Recent years of the Annual Social and Economic Survey similarly have 25% or higher missing data rates for earnings. Other variables such as race, gender and education are largely reported by survey respondents, and have missingness rates below 2%.

The missing data literature defines two potential mechanism for missingness. Missing completely at random (MCAR), the strongest assumption about the missingness process, implies that $\Pr[M = 1|Y, X] = \Pr[M = 1]$, where M is an indicator for Y being missing (Seaman et al (2013)). The second assumption is missing at random (MAR) where $\Pr[M = 1|Y, X] = \Pr[M = 1|X]$. Here, missingness may be related to X values, but not Y. If data are MCAR then $E[Y|M] = E[Y]$, If data are MAR then $E[Y|M, X] = E[Y|X]$ since Y is independent of M conditionally on X.

I will assume MCAR when considering inference in estimation of means, and MAR in considering inference in estimation of the regression. The assumption of MCAR is clearly too strong, and is obviously rejected by the literature. Indeed, when MCAR fails, and MAR holds, imputations can and do reduce or eliminate bias in estimation

of the mean from sample selection on X. I use MCAR below to simplify the results and focus on the issue of inference. Moreover, the simple model provides a framework for understanding the issue in the MAR case. The assumption of MAR however, is crucial to the Census hot deck and all imputation approaches (see Little and Ruben, 2014). Unfortunately it is seldom investigated and the few cases that have (see for example Lillard, Smith, and Welch, 1988; Bollinger and Hirsch, 2013; Bollinger et al, 2019) find at least some evidence against this assumption. The focus here, however, is not on the appropriateness of MAR, but rather under the assumptions made by Census, how do imputations impact sampling variance, inference, and efficiency. The reason often cited by researchers for using imputations is to improve efficiency by including additional observations.

3 Hot Deck Imputations

The hot deck imputation method of Census is a highly sophisticated approach to imputing individual missing variables. Census also has a "whole impute" hot deck approach for individuals who do not respond to the Annual Social and Economic Survey supplement. I focus here on earnings since the imputation rate is very high for this variable and earnings regressions are common. A detailed explanation of the hot deck procedure can be found in Hokayem, Raghunathan and Rothbaum (2020). Individuals who are employed but do not respond to the earnings questions are classified using sixteen key variables known to be correlated with earnings and or missingness of earnings. The variables are broken into multiple categories, and the intersection of these provide over 3.8 million cells or decks. In the 1970's when the approach was developed (with fewer variables and categories), data were processed on punch cards. When an observation came in with a completed response for earnings, after processing that observation, the punch card was placed in a the deck of punch cards corresponding to its appropriate categories. When an observation arrived for processing which was missing the earnings the top card on the "hot deck" for

that category was drawn and replaced the actual missing observation. Today the processing is completely electronic but the basic idea still holds. A missing value is replaced by a random draw from the bin of observed incomes defined by the sixteen variables and their categories. In the case of the ASEC, when a match cannot be found, cells are collapsed (see Hokayem, Raghunathan and Rothbaum, 2020) until a match is found. In essence, this is a highly sophisticated, non-parametric model which generates an imputation with the same mean and variance, conditional on a set of X's, as the missing observation.

To model the hot deck I consider two less complicated approaches. In the first approach, which I will call the simple hot deck, I assume MCAR and assume that an imputation is a random draw from the complete case sample. One perspective on this would be to consider all the data in a single cell of the actual hot deck approach. In the second approach, which I'll call the match hot deck, I consider a single X (presumably categorical), assume MAR (conditional on that X) for Y, and assume an exact match (based on X) can always be found. While this over-simplifies a sophisticated approach, it preserves two key elements of the hot deck procedure: imputations are randomly drawn observations from the complete case sample and the mean and variance of the original distribution of Y is preserved. Further, it implies (see appendix B) that all estimators below are unbiased under the hot deck procedure. The focus here is on sampling error and efficiency.

Throughout the paper, I will assume that a random sample from the population is drawn with n observations. Two variables are intended to be measured: Y_i and X_i (although X plays no role in some sections). For n_1 observations both variables are observed: $\{Y_i, X_i\}_{i=1}^{n_1}$. For n_2 observations only X_j is observed $\{., X_j\}_{j=1}^{n_2}$. By random sampling, the original observations are independent. Let Y_j be imputed values drawn either through the simple hot deck or the match hot deck. Without loss of generality, we will assume that $n_2 < n_1$ and that hot deck draws are done without replacement so that no donor appears twice. If the match hot deck is

employed, the match is perfect so that X_i from the donor = X_j from the recipient.

3.1 Sample Mean Case

The simple sample mean case is a useful departure point. Estimators and their sampling properties are straightforward. It is also quite similar to the regression case below, as it is often forgotten that the sample mean is simply the regression on only an intercept. Throughout this section I maintain four assumptions:

$$\begin{aligned} A1 &: E[Y] = \mu \\ A2 &: V(Y) = \sigma^2 \\ A3 &: MCAR \\ A4 &: \textit{simple hot deck} \end{aligned}$$

Let

$$\widetilde{Y}^* = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

be the ideal sample mean, where no observations are missing. Recall that under random sampling, A1, and A2, the ideal sample mean is unbiased and has a sampling variance of $\frac{\sigma^2}{n}$. Let

$$\widehat{Y} = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_i$$

be the complete case mean. Under MCAR and random sampling, the complete case mean is unbiased and has sampling variance $\frac{\sigma^2}{n_1}$. Let

$$\widehat{\widehat{Y}} = \frac{1}{n}\left(\sum_{i=1}^{n_1} Y_i + \sum_{j=1}^{n_2} Y_j\right),$$

be the imputed mean, where the Y_j are the imputed values. The key insight here is that any Y_j used an imputation is a repeat of a Y_i from the observed sample. The original sampling frame is random sampling, resulting in independence, but this is subverted by reusing observations. The imputed mean is also unbiased (see appendix

B) under these assumptions, but it is important here to point out that MCAR is crucial.

Consider the variance of our imputed mean

$$V\left(\widehat{\widehat{Y}}\right) = \left(\frac{1}{n^2}\right)\left(V\left(\sum_{i=1}^{n_1} Y_i\right) + V\left(\sum_{j=1}^{n_2} Y_j\right) + 2Cov\left(\sum_{i=1}^{n_1} Y_i, \sum_{j=1}^{n_2} Y_j\right)\right). \quad (1)$$

The covariance term arises because of the resampling from the original data. The assumption that the hot deck uses no donor twice is relied upon, but would only add a second covariance term. With the assumptions A2, A4, and MCAR this can be expressed in two convenient ways (see appendix B):

$$V\left(\widehat{\widehat{Y}}\right) = \left(\frac{\sigma^2}{n}\right)\left(1 + \frac{2n_2}{n}\right) = \left(\frac{\sigma^2}{n_1}\right)\left(1 + \frac{n_2}{n}\left(\frac{n_1 - n_2}{n}\right)\right). \quad (2)$$

There are two important comparisons here. First note that the sampling variance is larger than the ideal sample mean variance $\frac{\sigma^2}{n} < \left(\frac{\sigma^2}{n}\right)\left(1 + \frac{2n_2}{n}\right)$. This implies that our imputed sample mean is less efficient than the ideal sample mean. This is well understood in the imputation literature, but when typical computer commands are used, the estimated standard errors are generated using the ideal sample mean variance expression which understates the true sampling variance of the imputed sample mean estimator. Thus inference on these means is invalid.

The second important comparison is to $\frac{\sigma^2}{n_1}$ the complete case sampling variance. The second expression in equation 2, and the assumption that $n_1 > n_2$ demonstrate that the sampling variance of the imputed sample mean is larger than the sampling variance which would be achieved in the complete case estimate. The intuition on this is actually quite simple. The imputed sample mean is the complete case sample mean averaged with a random subsample of the complete cases. Thus variation increases in the same way that adding two correlated random variables together increases variation. If the complete case average is above (below) the true mean, then the additional imputed mean is likely to be above (below) the average as well, increasing sampling swings.

There are two interesting extreme cases: if $n_2 = 0$, then including imputations (none) results in an estimate that is equivalent to simply using the observed sample. The other extreme is when $n_1 = n_2$ where again, the resulting sample variance expression is equal to that achieved by the observed sample alone. Under the assumptions here this second case results in every actual observation being repeated as an imputation exactly once. This second case highlights a key insight of this note: you cannot improve your sample precision by using the data more than once. While the details of other settings will be slightly more complicated, this key insight drives all results.

3.2 Regression Setting

A regression setting allows the more realistic MAR assumption to be used. Throughout this section I assume the match hot deck is performed in an "ideal" manner: each draw of Y comes from a perfect match to X_j. Bollinger and Hirsch (2006) established that in the absence of perfect matching, imputations cause bias in all coefficients. In addition to random sampling, the following assumptions hold:

$$\begin{aligned} A5 &: Y_i = \alpha + \beta X_i + \varepsilon_i \\ A6 &: E\left[\varepsilon_i | X_i\right] = 0 \\ A7 &: V\left(\varepsilon_i | X_i\right) = \sigma^2 \\ A8 &: MAR \\ A9 &: \textit{match hot deck.} \end{aligned}$$

For simplicity, I will treat the $X's$ as non-stochastic. Alternatively, one could derive similar results conditional on the realized $X's$ or apply asymptotic results. The main point follows regardless. I focus on the regression slope from an ordinary least squares regression. Similar to the sample mean above, define

$$\widetilde{b}^* = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}$$

to be the ideal slope where no observations are missing. This is unbiased and has a sampling variance of

$$V\left(\widetilde{b}^{*}\right)=\frac{\sigma^{2}}{\sum_{i=1}^{n}\left(X_{i}-\overline{X}\right)^{2}}. \tag{3}$$

Let

$$\widehat{b}=\frac{\sum_{i=1}^{n_1}\left(X_{i}-\overline{X}_{1}\right)\left(Y_{i}-\overline{Y}_{1}\right)}{\sum_{i=1}^{n_1}\left(X_{i}-\overline{X}_{1}\right)^{2}}$$

be the complete case slope. Note that $\overline{X}_1$, the sample mean from the complete case sample may be very different than $\overline{X}$, the sample mean from the full sample. Both are observable to the researcher. Similarly, $\overline{Y}_1$ and $\overline{Y}$ may differ as well (in contrast to the section above). The MAR assumption ensures that

$$E\left[Y_{i}|X_{i},M_{i}=0\right]=E\left[Y_{i}|X_{i}\right]=\alpha+\beta X_{i}.$$

Thus the complete case slope is unbiased and its sampling variance is:

$$V\left(\widehat{b}\right)=\frac{\sigma^{2}}{\sum_{i=1}^{n_1}\left(X_{i}-\overline{X}_{1}\right)^{2}}. \tag{4}$$

The imputed slope coefficient will be

$$\widehat{\widehat{b}}=\frac{\sum_{i}\left(X_{i}-\overline{X}\right)\left(Y_{i}-\overline{Y}\right)+\sum_{j}\left(X_{j}-\overline{X}\right)\left(Y_{j}-\overline{Y}\right)}{\sum_{i}\left(X_{i}-\overline{X}\right)^{2}+\sum_{j}\left(X_{j}-\overline{X}\right)^{2}}.$$

The structure of $\widehat{\widehat{b}}$ is quite similar to that of the imputed sample mean above: the two terms in the numerator represent the averages for the observed (indexed by i) and the imputed (indexed by j). The key result hinges again on the fact that Y_j is perfectly correlated with its donor Y_i. Since the match is perfect, $X_j = X_i$ as well. The sampling variance of the imputed slope is

$$\begin{aligned} V\left(\widehat{\widehat{b}}\right) &= \frac{\sigma^{2}}{\left(\sum_{i}\left(X_{i}-\overline{X}\right)^{2}+\sum_{j}\left(X_{j}-\overline{X}\right)^{2}\right)}* \\ &\quad \left(1+\frac{2\sum_{j}\left(X_{j}-\overline{X}\right)^{2}}{\left(\sum_{i}\left(X_{i}-\overline{X}\right)^{2}+\sum_{j}\left(X_{j}-\overline{X}\right)^{2}\right)}\right). \end{aligned} \tag{5}$$

As with the simple mean case, the sampling variance with the hot deck imputations is larger than the ideal estimator sample variance. This expression is structurally identical to the one in the sample mean case: simply replace $\sum_i \left(X_i - \overline{X}\right)^2$ with n_1 and $\sum_j \left(X_j - \overline{X}\right)^2$ with n_2 and the result from the sample mean case emerges. The role of the terms like $\sum_i \left(X_i - \overline{X}\right)^2$ is similar to that of n_1 and n_2 as these sums rise with n in a similar way. The key result is that the usual formula for the sampling variance computed by statistics packages is wrong, and understates the standard errors of the actual estimate when imputations are included.

Comparison to the complete case estimator is more challenging under MAR, because $\overline{X}$ and $\overline{X}_1$ may differ. The variance of the imputation estimator can be expressed in terms of the variance of the complete case estimator:

$$V\left(\widehat{\overline{b}}\right) = \frac{\sigma^2}{\sum_{i=1}^{n_1}\left(X_i - \overline{X}_1\right)^2} * \left(1 + \left(\frac{\sum_j \left(X_j - \overline{X}\right)^2}{\sum_i \left(X_i - \overline{X}\right)^2 + \sum_j \left(X_j - \overline{X}\right)^2}\right) * \left(\frac{\left(\sum_i \left(X_i - \overline{X}_1\right)^2 - \sum_j \left(X_j - \overline{X}_2\right)^2\right) - \left(\frac{n_1^2 n_2}{n^2}\left(\overline{X}_1 - \overline{X}_2\right)^2\right)}{\sum_i \left(X_i - \overline{X}\right)^2 + \sum_j \left(X_j - \overline{X}\right)^2}\right)\right). \tag{6}$$

It is less clear here whether the term multiplying the variance of the complete case estimator is larger or smaller than 1. The term $\left(\frac{\sum_j \left(X_j - \overline{X}\right)^2}{\sum_i \left(X_i - \overline{X}\right)^2 + \sum_j \left(X_j - \overline{X}\right)^2}\right)$ is positive. The term $\left(\sum_i \left(X_i - \overline{X}_1\right)^2 - \sum_j \left(X_j - \overline{X}_2\right)^2\right)$ is similar to the difference $n_1 - n_2$ in that the two sums each have that many terms. However, it also involves the difference in variation of X between the observations with observed Y and missing Y. The term $\frac{n_1^2 n_2}{n^2}\left(\overline{X}_1 - \overline{X}_2\right)^2$ measures the squared difference in means.

Consider conditions on whether the term $\left(\sum_i \left(X_i - \overline{X}_1\right)^2 - \sum_j \left(X_j - \overline{X}_2\right)^2\right)$ would be positive (if it is negative, then certainly the imputation variance is less than the complete case variance). Simple algebra provides:

$$\frac{S_1^2}{S_2^2} > \frac{n_2}{n_1},$$

where S_1^2 is the sample variation in X for observations where Y is observed, while S_2^2 is the sample variation in X when Y is missing. Considering the case of the ASEC,

where the proportion of missing earnings is approximately 25%, the ratio $\frac{n_2}{n_1}$ would be 1/3. Thus, the term is positive provided that S_2^2 is no more than three times S_1^2. In an MCAR world, the term $\left(\sum_i \left(X_i - \overline{X}_1\right)^2 - \sum_j \left(X_j - \overline{X}_2\right)^2\right)$ will be dominated by the relative sample sizes, as the sample variances of the two groups will be quite close. Similarly the difference in means should be small, and the second term will be small.

The cost of not using the observations where Y is missing is determined by the loss in variation in X from the missingness process. OLS slope estimators have smaller sampling variance when the explanatory variables have larger variation. If missingness in Y reduces variation in X either through a reduction in variation around the sample mean, or by a reduction in variation due to a shift in the sample mean, the imputation estimator could result in improved precision of the estimate relative to the complete case estimate.

While the comparison of this case to the complete case estimator is less clear as to whether imputations can improve the estimation of linear model, the important result is that using hot deck type imputations in estimation requires a different expression for the standard errors than is typically used. The gains are smaller than one may think based purely on sample sizes. The cost of mismatch, as outlined in Bollinger and Hirsch (2006) is high.

4 Monte Carlo Results

Analytic results comparing the complete case analysis to the imputation estimator are clear for MCAR, but less clear for MAR. To investigate this, I present the results of two simulations. The first mimics the ideal set-up of the regression case, where the X variable is a simple binary regressor and is matched perfectly in the imputation process. The imputations are drawn randomly without replacement (simulation results with replacement are qualitatively similar and available upon request). The results for this include a MCAR case and cases with missing Y rates differing across

X values. The second simulation uses a continuous X variable and imputations are drawn from a categorical assignment based on the continuous X. This case is similar to how imputations in the CPS are constructed where age is divided into six categories and education is divided into three categories. The missing rate differs across the values of X. In both simulations, the results highlight that as long as X is included in the regression, even cases where missing is highly related to the X variable do not bias regression results. However, imputations may lead to bias with match error (Bollinger and Hirsch, 2006). Imputations must be done in the context of the model to be estimated (Little and Rubin, 2014; Hokayem, Raghunathan, Rothbaum, 2020). The imputation estimators were less efficient than the complete case. Using imputations often leads to less efficiency and higher true standard errors, even when match bias is not an issue.

4.1 Ideal Case

The data generating process for this case relies on a binary X variable with probability of being 1 equal to 0.40. The dependent variable $Y = 1+X+u$, where u is a standard normal variable and is generated independent of X. The full sample size is $n = 1000$. Whether the Y value is missing is determined by

$$m = 1[v + d * X < pm + d * px]$$

where v is a uniform random variable, d is the differential rate for missing data, pm is the overall rate for missing data and px is the probability $X = 1$. For example, with $d = 0$ and $pm = .25$, 25% of the Y variables will be set as missing, and the probability of missing will be the same regardless of the value of X. If $d = .1$, then if $X = 1$, the rate at which Y will be missing is $pm + d * px - d = .25 + .1 * .4 - .1 = .19$. While for $X = 0$ cases, the rate for missing Y will be 0.29: I use four differential rates: $0, .1 * pm,\ 0.25 * pm,\ 0.5 * pm$. I chose six different overall missing rates for pm: 0.05, 0.1, 0.15, 0.2, 0.25, 0.3. I provide three estimates of the slope coefficients: the ideal slope $\left(\widetilde{b^*}\right)$, the complete case estimator $\left(\widehat{b}\right)$, and the imputed estimator

$\left(\widehat{\widehat{b}}\right)$. I report three statistics for each estimator: the average slope, the variance of the slope, and the mean squared error of the slope from the ideal of 1. Five hundred repetitions are used.

Panel A of Table is the case where the missing rate is the same across the two values of X. The rows of panel A present the six different rates for missing Y data. The first three columns provide the average estimate of the slope coefficient. All three estimates are unbiased and consistent, and this is supported. The second three columns are the variance, which should rise across the three columns. The first column is the ideal estimator when no data are missing. As expected, the variance of the ideal should not change with the missingness rate. The variance for the complete case rises as the rate of missing data rises reflecting the declining sample size (at 5% the expected sample size is 950, at 30% it is 700). A similar pattern in the column where imputations are used highlights the key result: higher rates of imputations resulting in higher variance. Down all rows, the variance in the imputations case is higher than the variance in the complete case and rises with the missingness rate. The final three columns present the mean squared error. The mean squared error rises for both the complete case and imputation case as the missing rate and the imputation column has the higher mean squared error.

Panel B imposes a small differential of 0.1 times the missing data rate. Thus in the first row of panel B, where the overall missing rate is 5%, the differential is .5% or half of 1%. When $X = 1$, the missing rate is 4.7% while when $X = 0$, the missing rate is 5.2%. In the last row of panel B, the missing rate for $X = 1$ is 28.2% while the missing rate for $X = 0$ is 31.2%. Panels C and D repeat this with a differential of.$25 * px$ and $.5 * px$. The patterns described above hold in these two panels as well: the sampling variance and mean squared error of the imputation estimator rises with the overall missing rate, and is always higher than the complete case estimator. The differential in missing data rates does not play a significant role in either determining the variance or the mean squared error. Estimates using the hot deck imputation are

less precise than those using the complete case.

4.2 Imperfect Match Case

I display a set of results from the imperfect matching case studied by Bollinger and Hirsch (2006). The use of bins to group variables such as age and education results in biased coefficients. A continuous variable X is generated as a standard normal random variable. As above $Y = 1+X+u$ where u is standard normal and independent of X. Imputation is based on k different quantiles of the values of X. An observation with missing Y is imputed from the observations with X in the same group. The regression specification uses the actual value of X. This simulates Census hot deck using categories for age or education while researchers use age and years of education.

In table 2 the average of the estimated slopes, the variance of the estimated slopes, and the mean squared error are all presented for the ideal, complete case and imputation estimates. I focus on a missing rate of 25% which is completely at random (similar to panel A of table one where the rate does not depend on X). I use five different levels of groups for the imputation process: 1, 2, 4, 6 and 10. The case of one group simulates the case studied by Hirsch and Schumacher (2004) where the variable is not conditioned on in the imputation procedure. Other results (including MAR) are similar.

There is no particular observable bias pattern in the estimated complete case slope. Even if missing data in Y depended on the X variable this would not change unless the model were misspecified. Unlike table 1, the estimated slopes from the imputed estimator are attenuated. The bias should be largest for $k = 1$ as this completely severs the relationship between X and Y for all imputed values. It may be that ordering of data induced some improvement here. The bias is largest for small groups and declines as the number of groups rises. The variation and mean squared error of the imputed estimates is higher than the variation of the complete case estimate.

5 Empirical Example

The CPS ASEC data are used extensively by economists investigating the determinants of earnings. Hirsch and Schumacher (2004) and Bollinger and Hirsch (2006) have investigated the bias due to imputations when explanatory variables not used in the imputation approach are included in estimation. Bollinger and Hirsch (2013) examined whether sample selection on unobservables led to bias when non-respondents were removed from the sample. Bollinger et al (2019) demonstrate that MAR does not hold.

The data here derive from the March 2018 ASEC. The sample is limited to male heads of household between 18 and 64 years old who work full time year round. In the ASEC data, individuals who do not provide a full response to the ASEC supplement (but are participating in the monthly core) have their entire record imputed (whole imputes). This introduces imputed X's and thus I exclude them from this analysis.

Table 3 presents summary statistics for three samples: the full sample, the complete case sample, and the imputed sample. The first two columns present the mean and standard deviation of the full sample (n=14,281), including observations where the dependent variable is imputed. The variables chosen are standard variables in a basic Mincer earnings model: Age, race and education. Imputed earnings are 23.4% of the sample. The third and fourth column are the 10,993 complete case observations. The fifth and sixth column are the observations where earnings is imputed. There are few differences in mean or standard deviation between the Complete Case sample and the Imputed Sample. The imputed sample is slightly older and has a larger spread of ages, has slightly higher proportions of minorities, and is generally less well educated. The relatively small differences will likely be overwhelmed by the sample size and we would expect the corrected variance of the full case estimator to be larger than the complete case estimator.

Table 4 presents the estimates from using the full sample and the complete case sample. The standard errors calculated by Stata's regress command are reported in

the column "old S.E." (second column) while the standard errors which adjust for the imputations are reported in the column labeled "new S.E." (third column). The complete case coefficients and their appropriate standard errors, are reported in the fourth and fifth column. The coefficients differ little between the two estimates. The standard errors ("old S.E.") which treat the imputations as independent draws from the population are the smallest. The properly adjusted standard errors for the imputed estimator are larger than the standard errors for the complete case estimator. The differences in mean and variance between the complete case and imputed sample do not dominate the covariance induced by re-using Y. The result is clear: the standard errors produced normally by statistics packages understate the sampling variance when imputations are included in the dependent variable and the complete case estimator is more efficient. The alleged loss in sample size or spread is not remedied by the inclusion of imputations.

6 Conclusion

Economists are fond of the phrase, "There is no such thing as a free lunch." One cannot improve sample precision by simply repeating the same observations, even if you call them "imputations." There are good and important reasons for considering the implications of missing data, and imputations can, in the case of MAR, reduce bias in estimates of means. This is the setting where advocates of imputation methods typically focus attention, and rightfully so. Even then standard errors should be adjusted. In situations where estimation of a regression or more complex model is needed, the approaches needed for consistent estimation and inference are far more complicated than the use of stock imputations provided by a statistical agency. At the very least, a sound understanding of the imputation approach and its interaction with the model to be estimated is required, and stock estimators for standard errors cannot be used. The Census Bureau does provide a set of a adjustments for sampling variation (see, U.S. Bureau of the Census, 2018). However it is not clear that this

actually adjusts fully for hot deck imputations.

Another approach is to use sampling weights based on the known covariates with missing data (see Wooldridge, 2003; Moffitt, Fitzgerald, and Gottschalk, 1999). Indeed the hot deck procedure is simply a reweighting approach where observations which are under-represented in the observed data are replicated to construct a data set which mimics a more representative sample. The advantage to weights is that most computer packages are capable of using weights and adjusting sampling variance estimates appropriately for those weights.

In a regression setting, if we believe that Y is missing at random, conditional on a set of known $X's$ (a necessary condition for imputations), then provided our regression model is correctly specified on those variables, complete case estimates are still unbiased and consistent: the resulting sample of Y is random conditional on the $X's$ which is a sufficient condition. Under these assumptions it is difficult to find conditions where the imputation estimator improves upon the complete case estimator. Hirsch and Schumacher (2004) and Bollinger and Hirsch (2006) have shown conditions where imputations under MAR can result in bias. Hokayem, Raghunathan and Rothbaum (2020) demonstrate an improved imputation approach which addresses some of the concerns raised in Bollinger and Hirsch (2006)

Many Census products and many researchers have used various types of imputations to address missing data. While there are conditions (see Little and Rubin, 2014) under which imputation may be used very effectively, in particular when missing data occurs on the right hand side of the regression equation, researchers should be cautious. The key assumption of missing at random is also called into question by recent research of Bollinger and Hirsch (2013) and Bollinger et al. (2018). If the missing at random assumption fails, the standard hot deck or other approaches will not work. Bollinger and Hirsch (2013) demonstrate the use of a sample selection correction approach, but this too requires strong assumptions.

References

[1] Bollinger, Christopher R. and Barry T. Hirsch (2006) "Match Bias in the Earnings Imputations in the Current Population Survey: The Case of Imperfect Matching" Journal of Labor Economics, vol. 24. no. 3, 483-520.

[2] Bollinger, Christopher R. and Barry T. Hirsch (2013) "Is Earnings Response Ignorable?" Review of Economics and Statistics, Vol. 95, no. 2, 407-416.

[3] Bollinger, Christopher R., Barry T. Hirsch, Charles Hokayem, James P. Ziliak (2019) "Trouble in the Tails? What we know about Earnings Nonresponse Thirty Years After Lillard, Smith and Welch" Journal of Political Economy, vol. 127, no. 5, pp. 2143-2185.

[4] Current Population Survey, (2018) Annual Social and Economic (ASEC) Supplement conducted by the Bureau of the Census for the Bureau of Labor Statistics. – Washington: U.S. Census Bureau.

[5] Dempster, A.P. and D.B. Rubin (1983) "Introduction" Incomplete Data in Sample Surveys (Volume 2): Theory and Bibliography (W.G. Madow, I. Olkin and D.B. Rubin, eds) New York: Academic Press.

[6] Hirsch, Barry T. and Edward Schumacher (2004) "Match Bias in Wage Gap Estimates due to Earnings Imputation" Journal of Labor Economics, vol. 22, 689-722.

[7] Hokayem, Charles, Rivellore Raghunathan and Jonathan Rothbaum (2020) "Match Bias or Nonignorable Nonresponse? Improved Imputation and Administrative Data in the CPS ASEC." forthcoming, Journal of Survey Statistics and Methodology.

[8] Lillard, Lee, James P. Smith and Finis Welch (1986) "What do we really know about wage? The importance of nonresponse and Census imputations" Journal of Political Economy, vol 94, no. 3, part 1, 489-506.

[9] Little, Roderick J.A. and Donald B. Rubin (2014) Statistical Analysis with Missing Data, Wiley Series in Probability and Statistics, John Wiley and Sons.

[10] Moffit, Robert, John Fitzgerald and Peter Gottschalk (1999) "Sample Attrition in Panel Data: The Role of Selection on Observables" Annales d'Économie et de Statistique, No. 55/56, Panel Data Econometrics, pp.129-152.

[11] Seaman, Shaun, John Galati, Dan Jackson and John Carlin (2013) "What is Meant by 'Missing at Random'?" Statistical Science, Vol. 28, No. 2, pp. 257-268.

[12] U.S. Bureau of the Census (2018) "Calculating Approximate Standard Errors and Confidence Intervals for Current Population Survey Estimates", Current Population Survey, Technical Documentation, https://www.bls.gov/cps/calculating-standard-errors-and-confidence-intervals.pdf.

[13] Wooldridge, Jeffrey M. (2002) "Inverse Probability Weighted M-estimators for Sample Selection, Attrition, and Stratification" Portuguese Economic Journal, Vol. 1, pp. 117-139.

[14] Wooldridge, Jeffrey M. (2007) "Inverse Probability Weighted Estimation for General Missing Data Problems", Journal of Econometrics, vol. 141, pp. 1281-1301.

Table 1: Ideal Imputation Simulations, n=1000, reps = 500, true b = 1, Probability X =1 is 0.40

Panel A: Prob X =1 is .4, differential is zero (MCAR)									
	Mean			Variance			Mean Squared Error		
Probability of Missing	Full Sample	Complete Case	With Imputations	Full Sample	Complete Case	With Imputations	Full Sample	Complete Case	With Imputations
0.05	1.002	1.002	1.003	0.00439	0.00455	0.00479	0.00438	0.00455	0.00479
0.1	0.997	0.998	0.997	0.00393	0.00434	0.00484	0.00393	0.00434	0.00484
0.15	1.000	1.000	0.999	0.00361	0.00435	0.00508	0.00360	0.00435	0.00507
0.2	1.007	1.006	1.009	0.00361	0.00456	0.00541	0.00365	0.00459	0.00548
0.25	1.003	1.003	1.002	0.00447	0.00607	0.00674	0.00447	0.00607	0.00673
0.3	1.000	1.001	0.997	0.00411	0.00632	0.00704	0.00410	0.00631	0.00703
Panel B: Missing Differential is .1									
0.05	1.000	1.000	1.000	0.00467	0.00507	0.00530	0.00466	0.00506	0.00529
0.1	1.001	1.001	0.999	0.00396	0.00450	0.00477	0.00395	0.00449	0.00476
0.15	1.000	1.001	1.000	0.00360	0.00412	0.00442	0.00360	0.00411	0.00441
0.2	1.004	1.004	1.008	0.00440	0.00546	0.00587	0.00440	0.00547	0.00592
0.25	1.003	1.005	1.006	0.00389	0.00523	0.00582	0.00390	0.00524	0.00584
0.3	1.003	1.004	1.006	0.00396	0.00551	0.00602	0.00396	0.00552	0.00604
Panel C: Missing Differential is .25									
0.05	0.995	0.994	0.994	0.00417	0.00454	0.00462	0.00418	0.00456	0.00464
0.1	0.995	0.996	0.995	0.00384	0.00446	0.00481	0.00386	0.00447	0.00482
0.15	0.998	0.999	0.998	0.00426	0.00493	0.00560	0.00426	0.00492	0.00559
0.2	1.000	1.000	1.000	0.00401	0.00488	0.00538	0.00400	0.00487	0.00537
0.25	0.999	0.998	0.998	0.00428	0.00559	0.00607	0.00428	0.00558	0.00606
0.3	1.004	1.002	1.001	0.00449	0.00632	0.00712	0.00449	0.00631	0.00711
Panel D: Missing Differential is .5									
0.05	1.002	1.002	1.001	0.00403	0.00408	0.00418	0.00403	0.00408	0.00417
0.1	0.995	0.995	0.996	0.00387	0.00439	0.00463	0.00389	0.00441	0.00463
0.15	1.003	1.003	1.003	0.00453	0.00491	0.00553	0.00453	0.00491	0.00552
0.2	1.001	0.999	0.998	0.00421	0.00500	0.00560	0.00420	0.00499	0.00559
0.25	1.001	1.000	1.004	0.00402	0.00500	0.00540	0.00401	0.00499	0.00541
0.3	1.002	1.002	1.002	0.00417	0.00616	0.00682	0.00417	0.00615	0.00682

Table 2: Partial Match Imputation

	Mean Slope			Variance Slope			Mean Squared Error		
Imputation Groups	Full Sample	Complete Case	With Imputations	Full Sample	Complete Case	With Imputations	Full Sample	Complete Case	With Imputations
1	1.001	1.002	0.949	0.00104	0.00133	0.00165	0.00104	0.00133	0.00421
2	0.999	0.999	0.906	0.00105	0.00141	0.00181	0.00105	0.00141	0.01056
4	1.000	1.000	0.965	0.00099	0.00128	0.00155	0.00099	0.00128	0.00280
6	0.999	0.998	0.977	0.00099	0.00132	0.00164	0.00099	0.00132	0.00218
10	1.001	0.999	0.988	0.00102	0.00135	0.00153	0.00102	0.00135	0.00166

Table 3: Sample Summary Statistics

	Full Sample		Complete Case		Imputed Earnings	
VARIABLES	mean	sd	mean	sd	mean	sd
Age	42.91	11.03	42.68	10.93	43.66	11.29
Log(earnings)	10.99	0.735	11.00	0.725	10.95	0.766
Black	0.101	0.301	0.0915	0.288	0.131	0.338
White	0.793	0.405	0.807	0.395	0.748	0.434
Asian/Hawiian/Pacific Islander	0.0865	0.281	0.0825	0.275	0.0998	0.300
Native American	0.0196	0.139	0.0193	0.138	0.0206	0.142
Elementary School	0.0225	0.148	0.0211	0.144	0.0272	0.163
Some High School	0.0438	0.205	0.0431	0.203	0.0463	0.210
High School Degree	0.409	0.492	0.402	0.490	0.435	0.496
Associates Degree	0.108	0.310	0.110	0.314	0.100	0.300
Bachelors Degree	0.260	0.439	0.264	0.441	0.249	0.432
Masters Degree	0.109	0.312	0.112	0.315	0.0995	0.299
Profession/Ph.D. Degree	0.0466	0.211	0.0476	0.213	0.0436	0.204
Imputed Earnings	0.234	0.424	0	0	1	0
N	14,281		10,933		3,348	

Source: Authors calculations from March 2018 CPS ASEC, Male Head of Household, between age 18 and 64, no whole imputes, full time, full year workers with positive earnings.

Table 4: Regression Comparisons

	Full Sample			Complete Case	
	Coef F.S.	Old S.E.	New S.E.	Coef C.C.	Std. Err.
Age	0.067	0.0038	0.0046	0.074	0.0043
Age squared (000's)	-0.664	0.0442	0.0530	-0.732	0.0500
Black	-0.219	0.0182	0.0224	-0.230	0.0212
Asian/Hawiian/Pac Isd.	-0.012	0.0197	0.0238	-0.018	0.0225
Native American	-0.114	0.0392	0.0468	-0.152	0.0442
Elementary School	-0.497	0.0371	0.0454	-0.504	0.0428
Some High School	-0.341	0.0272	0.0326	-0.332	0.0307
Associates Degree	0.158	0.0185	0.0218	0.162	0.0206
Bachelors Degree	0.432	0.0137	0.0162	0.425	0.0153
Masters Degree	0.615	0.0187	0.0220	0.618	0.0207
Profession/Ph.D.	0.864	0.0267	0.0314	0.855	0.0297
_cons	9.253	0.0798	0.0954	9.078	0.0899

Source: Authors calculations from March 2018 CPS ASEC, Male Head of Household, between age 18 and 64, no whole imputes, full time, full year workers with positive earnings

Appendix B

Claim 1 *The sample mean using imputations is unbiased under random sampling and assumptions A1 through A4..*

$$E\left[\overline{Y}\right] = E\left[\frac{1}{n}\left(\sum_{i=1}^{n_1} Y_i + \sum_{j=1}^{n_2} Y_j\right)\right]$$
$$= \frac{1}{n}\left(\sum_{i=1}^{n_1} E\left[Y_i\right] + \sum_{j=1}^{n_2} E\left[Y_j\right]\right)$$

Simple random sampling, MCAR and assumption A1 imply that $E\left[Y_i\right] = \mu$. The assumption that the simple hot deck is a random draw, combined with random sampling, MCAR and assumption 1 also insure that $E\left[Y_j\right] = \mu$. Unbiasedness follows in the usual manner.

Claim 2 $V\left(\overline{Y}\right) = \left(\frac{\sigma^2}{n}\right)\left(1 + \frac{2n_2}{n}\right)$

As noted in the text

$$V\left(\widehat{\overline{Y}}\right) = V\left(\frac{1}{n}\left(\sum_{i=1}^{n_1} Y_i + \sum_{j=1}^{n_2} Y_j\right)\right).$$

Assuming that there are no double donors,

$$= \left(\frac{1}{n^2}\right)\left(V\left(\sum_{i=1}^{n_1} Y_i\right) + V\left(\sum_{j=1}^{n_2} Y_j\right) + 2Cov\left(\sum_{i=1}^{n_1} Y_i, \sum_{j=1}^{n_2} Y_j\right)\right).$$

With random sampling, MCAR, and random independent draws from the donors.

$$= \left(\frac{1}{n^2}\right)\left(\sum_{i=1}^{n_1} V\left(Y_i\right) + \sum_{j=1}^{n_2} V(Y_j) + 2Cov\left(\sum_{i=1}^{n_1} Y_i, \sum_{j=1}^{n_2} Y_j\right)\right).$$

With single donors,

$$Cov\left(\sum_{i=1}^{n_1} Y_i, \sum_{j=1}^{n_2} Y_j\right) = \sum_{j=1}^{n_2} V\left(Y_j\right)$$

since each Y_j corresponds to a single Y_i. Thus

$$V\left(\widehat{\overline{Y}}\right) = \left(\frac{1}{n^2}\right)\left(\sum_{i=1}^{n_1} V\left(Y_i\right) + \sum_{j=1}^{n_2} V(Y_j) + 2\sum_{j=1}^{n_2} V\left(Y_j\right)\right).$$

By assumption A2,

$$\begin{aligned} &= \left(\frac{1}{n^2}\right)\left(n_1\sigma^2 + n_2\sigma^2 + 2n_2\sigma^2\right) = \left(\frac{\sigma^2}{n^2}\right)(n + 2n_2) \\ &= \left(\frac{\sigma^2}{n}\right)\left(1 + 2\frac{n_2}{n}\right). \end{aligned}$$

Claim 3 *Under assumptions A5 through A9, and non-stochastic $X's$:* $V\left(\widehat{\widehat{b}}\right) = \frac{\sigma^2}{\left(\sum_i (X_i - \overline{X})^2 + \sum_j (X_j - \overline{X})^2\right)}\left(1 + \frac{2\sum_j (X_j - \overline{X})^2}{\left(\sum_i (X_i - \overline{X})^2 + \sum_j (X_j - \overline{X})^2\right)}\right)$

Consider $\varepsilon_j = Y_j - \alpha - X_j\beta$. Recalling that Y_j is an imputation and that exact matching on X_j is assumed, as in the expressions in the simple mean case, ε_j is an imputed ε and is drawn from the $\varepsilon_i's$ from the observed sample. (Note that if matching were not perfect, $\varepsilon_j = \varepsilon_i + (X_j - X_i)\beta$, the key part of the proof would hold: ε_j would still be a draw from the $\varepsilon_i s$). Then the approach here is nearly identical to the approach in the simple mean case above:

$$V\left(\widehat{\widehat{b}}\right) = V\left(\frac{\sum_i \left(X_i - \overline{X}\right)\varepsilon_i + \sum_j \left(X_j - \overline{X}\right)\varepsilon_j}{\sum_i \left(X_i - \overline{X}\right)^2 + \sum_j \left(X_j - \overline{X}\right)^2}\right) = \frac{V\left(\sum_i \left(X_i - \overline{X}\right)\varepsilon_i + \sum_j \left(X_j - \overline{X}\right)\varepsilon_j\right)}{\left(\sum_i \left(X_i - \overline{X}\right)^2 + \sum_j \left(X_j - \overline{X}\right)^2\right)^2}$$

$$= \frac{V\left(\sum_i \left(X_i - \overline{X}\right)\varepsilon_i\right) + V\left(\sum_j \left(X_j - \overline{X}\right)\varepsilon_j\right) + 2Cov\left(\sum_i \left(X_i - \overline{X}\right)\varepsilon_i, \sum_j \left(X_j - \overline{X}\right)\varepsilon_j\right)}{\left(\sum_i \left(X_i - \overline{X}\right)^2 + \sum_j \left(X_j - \overline{X}\right)^2\right)^2}$$

$$= \frac{\sigma^2\left(\sum_i \left(X_i - \overline{X}\right)^2 + \sum_j \left(X_j - \overline{X}\right)^2\right)}{\left(\sum_i \left(X_i - \overline{X}\right)^2 + \sum_j \left(X_j - \overline{X}\right)^2\right)^2} + 2\frac{\sigma^2\sum_j \left(X_j - \overline{X}\right)^2}{\left(\sum_i \left(X_i - \overline{X}\right)^2 + \sum_j \left(X_j - \overline{X}\right)^2\right)^2}$$

$$= \frac{\sigma^2}{\left(\sum_i \left(X_i - \overline{X}\right)^2 + \sum_j \left(X_j - \overline{X}\right)^2\right)} + 2\frac{\sigma^2\sum_j \left(X_j - \overline{X}\right)^2}{\left(\sum_i \left(X_i - \overline{X}\right)^2 + \sum_j \left(X_j - \overline{X}\right)^2\right)^2}$$

$$= .\frac{\sigma^2}{\left(\sum_i \left(X_i - \overline{X}\right)^2 + \sum_j \left(X_j - \overline{X}\right)^2\right)}\left(1 + \frac{2\sum_j \left(X_j - \overline{X}\right)^2}{\left(\sum_i \left(X_i - \overline{X}\right)^2 + \sum_j \left(X_j - \overline{X}\right)^2\right)}\right).$$

This result is straightforward to extend to a multiple regression

$$V\left(\widehat{\widehat{b}}\right) = \sigma^2\left(X'X\right)^{-1}\left(I + 2\left(X_j^{'}X_j\right)\left(X'X\right)^{-1}\right).$$