Self-Employment Status: Imputations, Implications, and Improvements

Jonathan Eggleston
U.S. Census Bureau

Mark A. Klee
U.S. Census Bureau

Robert Munk[†]
U.S. Census Bureau

March 2022
SIPP Working Paper Number 303
SEHSD Working Paper Number 2022-06

Abstract: We show that reports of self-employment status in the Survey of Income and Program Participation are not missing at random, raising concerns that the hot deck imputation technique for self-employment status is biased. This critique likely applies to all other U.S. Census Bureau household surveys, which currently use a hot deck to impute self-employment status. We find that combining the Sequential Regression Multiple Imputation (SRMI) technique with administrative employment records improves average imputation accuracy relative to the hot deck. SRMI preserves some correlations that hot deck imputation does not. But despite these improvements the data imputed by this new model falls short in matching all the correlations present in reported data. Analysts must, therefore, carefully decide how to handle respondents with imputed self-employment status.

Keywords: Self-employment Imputation, Survey Data, Tax Data, Hot Deck, Incorporation Status Imputation

JEL Codes: L26, C81, C83, J49, L22

Where relevant, all comparative statements in this paper have undergone statistical testing, and, unless otherwise noted, all comparisons are statistically significant at the 10 percent level. All data are subject to error arising from a variety of sources, including sampling error, non-sampling error, model error, and any other sources of error. For further information on the Survey of Income and Program Participation statistical standards and accuracy see:https://www2.census.gov/programs-surveys/sipp/tech-documentation/source-accuracy-statements/2014/sipp-2014-source-and-accuracy-statement.pdf

[†]Corresponding author: robert.o.munk@census.gov

I Introduction

Since at least the 1980s hot deck imputation has been the U.S. Census Bureau's default imputation model for self-employment status.[1] Hot deck imputation—which was first implemented by the U.S. Census Bureau in the 1960s (Ono and Miller, 1969; Cantwell et al, 2005) and derives its name from feeding a deck of punch cards into a mainframe computer (Sage Publications, 2008)—matches nonrespondents to respondent donors who share similar characteristics. For example, a hot deck might assign the self-employment status of a twentysomething, college educated, male respondent to a twentysomething, college educated, male nonrespondent.

Hot deck's computational feasibility meant that it was likely the best imputation technique when it was first implemented, but later developments have enabled methods that were literally impossible or practically infeasible in prior decades. With the number of transistors on a computer chip doubling roughly every two years (Theis and Wong, 2017), computers became exponentially more powerful and capable of more sophisticated techniques. In addition, this increased computational power along with the increased availability of administrative data enabled the linking of surveys to administrative data and the inclusion of administrative data into imputations. A deeper understanding of the limitations of hot deck imputation has also allowed survey administrators to address hot deck's known weaknesses. Combined these developments have opened the door to higher quality imputations. Developments that the U.S. Census Bureau has leveraged to better impute program participation and employment status in the Survey of Income and Program Participaiton (Benedetto et al. 2015, Giefer et al. 2015).

Until recently the U.S. Census Bureau had yet to apply these developments to the imputation of self-employment status for any of its household surveys: the American Community Survey, the Current Population Survey, and the Survey of Income and Program Participation.[2]

---

[1] Starting in 1980 hot deck imputation was used to impute self-employment status for the 1940 Census Public Use Microdata sample. Self-employment status may have been imputed by hot deck as early as the 1962 Current Population Survey, when labor income was first imputed by hot deck, but, despite extensive searching, we have been unable to find documentation to verify this.

[2]The Survey of Income and Program Participation uses a variant of a last observation carried forward hot deck. The data set is sorted on the variables in the hot deck (*e.g.*, sex, age, and race) as well as geography, and then the last observation with a reported value is carried forward and allocated to the missing value. There are two modifications employed by the U.S. Census Bureau: (1) the number of times an observation can be carried forward is limited, so no donor can contribute to a disproportionate number of missing cases; (2) if no donor is available, then the variable is assigned a cold-deck value (*i.e.*, a predetermined value selected by the U.S. Census Bureau). For more

Starting with the 2018 survey year, the Survey of Income and Program Participation (SIPP) no longer imputes self-employment status by hot deck. Instead self-employment status, along with incorporation status, is imputed with a regression based approach, Sequential Regression Multivariate Imputation, or SRMI. [3]

This analysis examines the quality of the SRMI imputation of self-employment and incorporation status in the SIPP and compares it to the quality of hot deck imputation. SRMI addresses two chief limitations of hot deck imputation. It enables the use of many more respondent characteristics in its imputations and allows for the easy inclusion of variables from administrative data. In particular, we include administrative measures of self-employment and firm size; two administrative variables that are correlated with survey self-employment and incorporation status.

There are three principal reasons why our analysis of the limits of hot deck imputation and benefits of SRMI matters for self-employment researchers.

First, estimates using observations with hot deck imputed self-employment and incorporation statuses are likely affected by match bias (Hirsch and Schumacher 2004, Bollinger and Hirsch 2006). In other words, the coefficient estimates of regressors not used to match donors to nonrespondents will be biased towards zero in regressions of self-employment status and self-employment transitions. Match bias occurs because covariates not included in the hot deck matching algorithm are not as strongly correlated with imputed values of the dependent variable as they are with reported values of the dependent variable. The correlations between the omitted match variables and the imputed dependent variable are entirely determined by the correlations between the omitted match variables and the variables in the matching algorithm. For example, if household wealth is not a variable used in the self-employment hot deck matching algorithm, the correlation between household wealth and imputed self-employment statuses will be determined by the correlation between household wealth and the matching variables, such as race and age. Match bias is a well-documented phenomenon within the

information see the SIPP 2014 Users' Guide: https://www2.census.gov/programs-surveys/sipp/tech-documentation/methodology/2014-SIPP-Panel-Users-Guide.pdf

[3] This imputation method is also commonly referred to as Sequential Regression *Multiple* Imputation. While Sequential Regression *Multivariate* Imputation produces only one implicate for variables with missing data, Sequential Regression *Multiple* Imputation produces more than one implicate for variables with missing data by taking repeated draws from a distribution of random noise. The additional implicates allow for estimates of standard errors that incorporate better the uncertainty introduced by imputation. SIPP public use data produces only one implicate for variables with missing data.

earnings literature. For example, Hirsch and Schumacher (2004) find that the union wage gap is substantially underestimated in Current Population Survey data because union status is not a matching criterion for Current Population Survey hot decked earnings. But in the self-employment literature the presence of match bias has not been documented and analysts regularly use imputed observations, likely introducing match bias.

Second, we examine whether self-employment data is missing at random. There are two common approaches to handling nonresponse in surveys: (1) retain provided imputations, treating nonrespondents identically to respondents and (2) forego imputation and explicitly drop all item nonrespondents in an attempt to avoid match bias. As we discussed above, the first approach can lead to match-bias, but the second approach may result in nonresponse bias if self-employment is not missing at random.[4] Nonresponse bias results from self-employment nonresponse being related to self-employment status itself. For example, individuals with non-traditional self-employment—such as contractors, gig economy workers, or day laborers—may be less likely to respond to questions about self-employment status out of confusion or an unwillingness to discuss their work, implying self-employment would suffer from nonresponse bias. Using administrative measures of self-employment status, we examine the relationship between self-employment nonresponse and administrative data to explore if self-employment suffers from item-level nonresponse bias. Our paper, thus, informs researchers and survey administrators about how two common biases—match and nonresponse—affect self-employment analyses, allowing for more informed decisions when handling missing self-employment data.

Third, increasing item nonresponse rates potentially increase match and nonresponse bias, increasing the importance of understanding how to handle missing data. Between 2000 and 2020, the imputation rate of the self-employment statuses roughly doubled in the Annual Social and Economic Supplement of the Current Population Survey (*i.e.,* the "March" CPS), increasing from ten to twenty-three percent (Figure 1). An increase that is consistent with the worrying trend of increasing unit nonresponse rates and item nonresponse rates for earnings and transfer income items (Bollinger et al, 2019; Meyer et al, 2015).

While the prospect of increasing match and nonresponse bias is worrying, improved imputation techniques combined with administrative data offer hope to reduce, or potentially

[4] Nonresponse bias is also commonly referred to as nonignorable nonresponse.

even eliminate, nonresponse and match bias. Regression based methods, such as SRMI, allow for the inclusion of far more covariates than hot deck methodologies, likely leading to reduced match bias. Administrative data also offers direct insight into the nature of a nonrespondent's employment relationship that can be incorporated into imputation models, which if sufficiently high quality may remove both match and nonresponse bias completely.

In our analysis, we take a pseudo-experimental approach and create a large sample of nonrespondents. Specifically, we impute self-employment and incorporation status to all respondents of the first wave of the 2014 SIPP panel. To avoid overfitting, we divide the sample into four equal groups and then impute each group using the other three groups as our sample of respondents. While this approach explicitly imposes a missing at random nonresponse for respondents, it allows us to examine match bias induced by both hot deck and SRMI imputation.

We compare four imputation methodologies: a Bayesian Bootstrap model—a hot deck look-alike, which is equivalent to hot deck imputation except that it assigns donors randomly rather than deterministically—with identical covariates to the current hot deck model; a Bayesian Bootstrap with administrative data indicators in addition to the covariates in the current hot deck model; an SRMI model with no administrative data; and an SRMI model with administrative data indicators.

Overall, we find that each model imputes high-quality aggregate statistics, but that the SRMI model with administrative data is the most likely to correctly impute a respondent's self-employment and incorporation status. While this is insufficient for concluding that SRMI with administrative data is a superior imputation methodology—assigning all respondents salary employment would "improve" upon SRMI with administrative data imputation based on this metric—it provides suggestive evidence of improved imputations.[5] To further evaluate, we run a logit model for the probability of being self-employed for five data samples: using the reported data and imputed data from each imputation methodology. We find that the SRMI model with administrative data matches the marginal effects of the reported data better than the model using our hot deck imputation look-alike does.

[5] The goal of imputation is to preserve the relationships between all variables, so that an analysis of imputed data would return the same results as if the data had been reported. Correct predictions may be consistent with this goal, but does *not* guarantee it is met. Assigning all jobs wage and salary employment better predicts a job's self-employment status than either SRMI imputation or hot deck imputation. But by assigning all jobs wage and salary employment, we would fail preserve the relationships between self-employment status and other variables; therefore, we would fail to achieve the goal of imputation.

Despite these improvements, the SRMI model poorly captures relationships in the reported data between self-employment status and some variables not included in the model. The marginal effects of these variables are biased towards zero, indicating match bias. This is most clearly seen in the marginal effect of having a self-employed partner on the probability of being self-employed. The marginal effect of having a self-employed partner is 73.8 percentage points in the reported data, while in the SRMI with administrative data imputations it is 27.1 percentage points and statistically indistinguishable from zero in the Bayesian Bootstrap model without administrative data (see Table 8). While newly discovered predictors of self-employment status can be added to the set of regressors to improve imputations, the feasible set of predictor variables is too large to capture every possible relationship with self-employment status.

Analysts might be hopeful that dropping nonrespondents would avoid match bias, but our analysis of nonresponse biases indicates that analysts are in a double bind: keeping imputations may subject estimates to match bias and dropping nonrespondents likely subjects estimates to nonresponse bias. We find that even after controlling for a comprehensive set of covariates, individuals without tax forms are more likely to have their self-employment status imputed, implying the missing at random assumption—which is necessary for hot deck imputation to yield unbiased predictions—is violated.

Therefore, while we conclude the new imputation technique constitutes an improvement over the current methodology, careful decisions are necessary when deciding how to handle respondents with imputed self-employment statuses. Analysts without access to administrative data might consider estimating their models with imputed responses, without imputed responses, and re-weighting respondents to account for response probability using survey variables. While none of these approaches would likely fully address match and nonresponse bias, the combination of approaches would allow researchers to assess their likely impact. Analysts with access to administrative data could also include administrative measures in their re-weighting approach. This approach would likely be the best means of addressing both biases, but our examinations here would not be sufficient to completely rule out nonresponse bias.

II Data Description

In this subsection, we discuss our three primary data sources: wave one of the 2014 Panel of the Survey of Income and Program Participation, the Social Security Administration's Detailed

Earnings Record, and the universe of Internal Revenue Service W-2s (W-2s). We then describe how they are linked.

The 2014 SIPP follows a nationally representative sample of households for up to four years. As with the preceding SIPP panels, the primary goals of the 2014 panel were to track short-term employment and income dynamics, household composition, and eligibility and participation in government assistance programs. The sample of households is drawn from the universe of the non-institutionalized U.S. population with an oversample of high-poverty areas. This oversampling enables more precise estimates of program participation but implies low-income households will be overrepresented in unweighted SIPP data.

The primary questions we utilize are the self-employment and incorporation status questions.[6] Field interviewers ask these questions for each job the respondent held (up to seven may be reported), so self-employment status can be determined at the job level. For each self-employed job, respondents are asked to report the incorporation status: unincorporated or incorporated.

The self-employment question in the SIPP is distinct from other surveys because it has three categories—wage and salary employment, self-employed business owner, and some other work arrangement—and directly equates business ownership and self-employment. The "some other work arrangement" category was added to SIPP in the 1996 panel to capture jobs that do not fit clearly in the employer or self-employed classification. These "other work arrangements" tend to be informal, impermanent, or irregular relationships with employers (*e.g.*, babysitting, odd jobs, and some freelancing).[7] The conflation of business ownership and self-employment is consistent with how the Internal Revenue Service views self-employed jobs. But practically speaking, many self-employed respondents do not view themselves as business owners and some wage-employed respondents are employed by their own business (Light and Munk, 2018).

[6] These are the ejb(n)_jborse and ejb(n)_incpb for n = 1 to 7 variables in the SIPP microdata.

[7] SIPP interviewers are instructed to use specific criteria to distinguish between self-employed business owners and workers in an "other work arrangement". A worker should be classified as self-employed if any of the following conditions are met machinery or equipment of substantial value is used in conducting the business; an office, store, or other place of business is maintained; or the business is advertised to the public. For example, a freelancer should be categorized as self-employed if any of these three conditions are met, and otherwise as a worker in an "other work arrangement". In practice, the decision between classifying a worker as self-employed and in an "other work arrangement" is likely influenced to a large extent by how the interviewer and the respondent interpret the question text and the criteria for business ownership.

Our primary administrative data source is the Detailed Earnings Record (DER) of the Master Earnings File from the Social Security Administration. The DER is a longitudinal history of annual earnings going back to 1978, which contains both administrative employee and self-employed earnings. For employees, the DER contains data from every W-2 a person received. For the self-employed, the DER contains person-level earnings from reported Form 1040, Schedule SEs (1040-SE). The SIPP is linked to the DER at the individual level using U.S. Census Bureau's Person Identification Validation System, as described in Wagner and Layne (2014). This procedure matches both survey data and administrative data to a master reference file. Individuals who are matched are given an identifier called a Protected Identification Key (PIK), which acts as an anonymized social security number that can be used to link administrative datasets and surveys. For 2014 SIPP Wave 1, 91.9 percent of individuals over 15 years old were assigned a PIK (Eggleston and Reeder 2018).

It is important to note that our survey and administrative measures of self-employment do not completely match. For example, unincorporated self-employed workers are only required to file a 1040-SE if their profits exceed $400, so individuals with low or negative profits will not show up in the DER. In addition, incorporated business owners receive a W-2 from the business they own—unless the business is an LLC and they choose to file as an unincorporated business owner. Therefore, incorporated self-employed workers will often show up as working for an employer in the administrative data despite claiming self-employment in the SIPP.

In part to address the imperfect match between the SIPP and the DER, we include an additional administrative variable correlated with self-employment: the number of employees that work at a firm. The variable is constructed from the universe of W-2s. For each W-2, we count the number of total W-2s with an identical employer identification number (EIN) and use this count as an approximation of the number of employees at a firm. The logic being that if a respondent works at a company with few or only one employee, then it is more likely they are self-employed. The methodology is imperfect because firms can have multiple EINs, and EINs with one W-2 can often be nannies or other types of household employees, but as we show in a later section the number of W-2s issued by an EIN is strongly correlated with survey self-employment.

III Methodology

III.A. A Description and Discussion of Hot Deck Imputation

Before describing the new method for imputing self-employment status, we first discuss how self-employment status was imputed by hot deck. A hot deck fills in missing values by copying to the nonrespondent values that were reported by a donor with an identical set of observable characteristics. For the 2014 Panel, donors and recipients were matched by sex, race, marital status, educational attainment, and age.

While prior research has established that hot deck imputation effectively replicates aggregate statistics (Andridge and Little, 2010), by necessity the number of determinants of self-employment included is limited due to the curse of dimensionality—the addition of matching variables exponentially increases the number of cells. The current self-employment hot deck has two-hundred and forty cells. The addition of one binary variable would increase the cell size to four-hundred and eighty. The addition of two binary variables would increase the cell size to nine-hundred and sixty. So each additional variable added substantially decreases the probability of having a sufficient number of donors for each cell. As a result, some determinants of self-employment status, such as foreign-born status and wealth, must be excluded from the set of observables used to match nonrespondents to donors, while others, such as education, must enter in coarser detail than properly explains self-employment.

III.B. A Description and Discussion of SRMI

Starting with the 2018 SIPP panel, Sequential Regression Multivariate Imputation (SRMI), as described by Raghunathan et al. (2001), was used to impute self-employment status and incorporation status. This, however, was not the first implementation of SRMI in the SIPP. SRMI made its SIPP debut in the 2014 panel and was used to impute a limited set of variables; for example, whether someone had a job.

Benedetto et al. (2015) describe the SRMI methodology as applied to SIPP. The model includes independent variables to predict missing values as either stratifiers or regressors. Each variable imputed by SRMI is regressed on a set of specified regressor variables. These regressions are run separately for each set of respondents having identical values of the categorical stratifier variables, with a model selection algorithm keeping only the regressor variables that are the most important predictors for each regression. Once regression coefficients are estimated, nonrespondents are classified according to the values of their stratifier variables and the relevant coefficient estimates are used to predict their missing values. Stratifiers thus

define the set of observables over which the regressors are assumed to have heterogeneous effects on the outcomes to be imputed.

A main benefit of this parametric approach is that it allows one to include many more explanatory variables in the imputation model. The inclusion of more explanatory variables preserves a wider range of correlations when predicting missing data, decreasing match bias. To do this, SRMI imputations proceed sequentially as a means of imputing multiple variables jointly. The benefit of sequential imputation is that it allows for other variables with missing data to be included as predictors. The imputation sequence iterates to include updated predictions of variables with missing data for predicting a variable of interest. For example, if someone has a missing value for both self-employment status and whether they are receiving Supplemental Nutrition Assistance Program (SNAP) benefits, then this method allows for self-employment to predict SNAP receipt, and vice versa. In the previous hot deck method, variables with missing values were completed sequentially without iterating this sequence. The sequential nature of the hot deck method implied that the set of variables available to match donors to nonrespondents depended heavily upon position in this sequence. This restriction severely limited the set of variables that could be imputed jointly, effectively preventing the joint imputation of variables with subjects as different as self-employment and SNAP receipt.

The other main improvement of this new approach is that we can include administrative variables as stratifiers or regressors, which should improve the quality of predicted values. While there are noteworthy discrepancies between administrative and survey measures of self-employment status (Abraham, Haltiwanger, Sandusky, et al, 2021), there are also strong correlations between the measures. Thus, the inclusion of administrative measures of self-employment likely increases imputation quality, even if some reconciliation among administrative and survey measures of self-employment is needed.

The administrative data sources described above are not the only administrative sources used by the SRMI imputation but they are the most relevant for self-employment and incorporation status. When SRMI imputations were introduced in the 2014 panel of the SIPP the U.S. Census Bureau leveraged multiple sources of administrative data. In addition to the DER, various social security benefit administrative data were used. These data were particularly relevant for predicting whether someone was receiving social security benefits. Therefore, these other administrative variables will be included among the set of regressors.

III.C. The Truth is Out There…But Usually Unobserved

One obstacle to evaluating the quality of imputed data is that the "true" value of the variable being imputed typically is unobserved. We address this challenge by restricting our sample to jobs with reported self-employment status and incorporation status. We randomly split our final analysis sample into quarters. For each quarter, we treat the reported self-employment status and incorporation status as missing, and we estimate the imputation model on the other three quarters. We then apply the model estimates to predict self-employment status and incorporation status for observations in the quarter with the counterfactually missing data. Consequently, for each observation in our final analysis sample, we observe both imputed values and the "true" reported values of self-employment status and incorporation status. Note that these reported values of self-employment status and incorporation status are subject to measurement error, and thus do not necessarily reflect respondents' objective status. Nevertheless, these reported values serve as a useful benchmark because imputation algorithms attempt to predict these observed values.

III.D. Sample Selection

Table 1 lists our sample selection criteria and enumerates how each condition influences our sample size. We begin with a full sample of about 40,000 jobs in wave 1 of the 2014 SIPP panel, representing calendar year 2013.[8] Most of our sample loss comes from dropping imputed values of self-employment status and incorporation status. Table 1 illustrates that the most common type of nonresponse occurs among individuals whose entire job was imputed, either because they ended their interviews before answering any employment questions or because they declined to provide any information about a job. These individuals are never asked questions about their self-employment status or their incorporation status on these jobs. Table 1 shows that it is less common for individuals to provide no response to these questions when they are asked. For regression analysis, we additionally drop individuals in a military industry or occupation and individuals living outside the 50 states and the District of Columbia.

III.E. Descriptions of Our Four Imputation Methodologies

[8] Since we perform unweighted analyses throughout this paper, sample counts are presented after applying rounding rules as articulated by the U.S. Census Bureau's Disclosure Review Board. For further details, see: https://www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/Rounding_Rules_Memo_v7.pdf.

SIPP's implemented SRMI methodology has two theoretical advantages over hot deck imputation: (1) SRMI includes administrative records among the set of explanatory variables and (2) SRMI can include many more predictor variables because it is not subject to the curse of dimensionality.[9] To gauge the relative importance of these two advantages, our evaluation of imputed data examines four methodologies.

1. Bayesian Bootstrap without administrative data: This method mimics the hot deck imputation previously employed by the SIPP. It is the baseline we use to examine the marginal improvements from adding administrative data and from using SRMI.The Bayesian Bootstrap has two noteworthy differences from the hot deck employed by the SIPP.
    a. The Bayesian Bootstrap characterizes workers according to the same variables used in the hot deck—sex, marital status, education, age, and race—but then *randomly* (as opposed to deterministically) assigns a self-employment status to nonrespondents from a donor who is identical along these dimensions.
    b. The Bayesian Bootstrap also differs from a hot deck in how it handles nonrespondents with no available donors. In the traditional hot deck if a donor cannot be found a cold deck value is assigned—a predetermined value assigned by Census analysts. Whereas a Bayesian Bootstrap uses a model selection algorithm to ensure that donors are matched to nonrespondents according to the best predictor variables by dropping less important predictor variables if no match can be found
2. Bayesian Bootstrap with administrative data: A Bayesian Bootstrap with identical predictors to the baseline Bootstrap, but with administrative data included. This method then allows us to understand the relative importance of administrative

[9] While administrative records are available for both respondents and nonrespondents who were assigned a PIK, administrative records are unavailable for individuals who were not assigned a PIK. To enable imputation of survey self-employment status for nonrespondents who were not assigned a PIK, SRMI also imputes the variables derived from administrative data for these individuals. Similarly, SRMI imputes the variables derived from administrative data to enable observations without a PIK but with reported self-employment status to contribute to the identification of SRMI coefficient estimates.

data by adding administrative records to the variables included in SIPP's prior hot deck.

3. SRMI without administrative data: This method is identical to the SRMI method used by the SIPP, but does not contain any administrative covariates. This model allows us to understand the relative importance of model-based imputation, which includes more observable explanatory variables than the hot deck. It also allows us to answer the question, how much does administrative data improve SRMI's imputation quality? A question that is both theoretical and practical. Because if Census did not receive administrative tax data in time for imputation purposes, SIPP's imputation of self-employment and incorporation status would use this model.
4. SRMI with administrative data: The model that implemented in the SIPP beginning with 2018 survey year. It includes a detailed set of survey variables and the administrative variables discussed above.

With each candidate imputation method, we create five implicates, that is impute a self-employment status for each job five times. While the SRMI methodology used in SIPP production only creates one implicate, the random variation across these implicates allows us to obtain standard errors for Tables 5 through 7. In Table 8, for our regression estimates, we use the first implicate only because the SIPP public use file only contains one implicate.

IV Results

IV.A. Does Administrative Self-Employment Predict Survey Self-Employment?

Table 2 lists the reported distribution of self-employment status and incorporation status among our final analysis sample at the person level. Approximately 11.5 percent of our sample reports at least some self-employment in the survey, with unincorporated self-employed jobs being more common than incorporated self-employed jobs.[10]

Consistent with Abraham, Haltiwanger, Hou et al. (2021), we find evidence of (1) broad disagreement between administrative and survey measures of self-employment, and (2) a

[10] Note that our population of interest is identical to our full analysis sample: SIPP respondents with reported self-employment and incorporation status, who were not in a military industry or occupation, and who did not live outside the 50 states and the District of Columbia. Consequently, inferences drawn in this paper are descriptive of the SIPP 2014 respondents and not necessarily representative of the U.S. population at large. All estimates in this paper are unweighted and do not account for the sample design.

substantial fraction of self-employed individuals who have neither a W-2 nor 1040-SE. For example, among individuals who reported only unincorporated self-employment in the 2014 SIPP, about 18 percent had wage and salary earnings in the DER and around 32 percent had no tax form in the DER. The disagreement evident throughout this table is both expected and worrying. It is consistent with survey and administrative measures differing conceptually—incorporated self-employed workers receive W-2s—but it is also consistent with reporting error—unincorporated self-employed workers with no other jobs should not receive W-2s. The lack of tax forms for the self-employed is also consistent with both differing concepts of self-employment between survey and administrative data and differential measurement error across these data sources. The lack of tax forms for a self-employed worker may indicate low or negative profits—recall unincorporated self-employed workers are only required to file a 1040-SE if their earnings exceed $400—or it may indicate the self-employed worker did not report their income to the IRS, a well-established pattern among the self-employed (Slemrod, 2007).[11]

Table 2 also reveals that the predictive power of administrative data likely differs by incorporation status for the self-employed. Among individuals who reported only incorporated self-employment in the SIPP, 20 percent had 1040-SE income compared to 50 percent of the unincorporated self-employed reporting 1040-SE income. Moreover, 50 percent of individuals reporting only incorporated self-employment had only W-2 income according to the DER, which is nearly five times the percent of individuals reporting only W-2 income by only unincorporated self-employed individuals.

Despite broad disagreement between survey and administrative measures of self-employment, Table 2 reveals that the source of administrative earnings is a potentially promising predictor of survey self-employment status. For instance, among individuals who only reported working for an employer in SIPP, about 91 percent displayed only W-2 income and only about 1 percent displayed only self-employment income in the DER. Similarly, among individuals who reported only unincorporated self-employment in the SIPP, about 18 percent received any W-2 income according to the DER and 50 percent received only self-employment income in the DER.

IV.B. Does Firm Size Predict Self-Employment Status?

[11] The sample for Table 2 also includes individuals who are not assigned a PIK. Because no tax forms can be assigned to them, they appear in the "Neither" column because they lack tax forms due to measurement error.

The universe of W-2s provide another potentially promising predictor of self-employment status based on administrative data: firm size. For sample members who received a W-2, we can identify the number of individuals who received a W-2 from this EIN. This measure only proxies for firm size because some firms issue W-2s under multiple EINs and we have no measure of which EINs belong to the same firm. Table 4 presents coefficient estimates from a job-level Ordinary Least Squares regression of a binary self-employment indicator on firm size indicators. These firm size categories are person-level variables, meaning that two firm size indicators will take a value of one for a person who received a W-2 from two employers of different firm size categories. Missing firm size indicates that an individual did not receive a W-2, implying they either filed a 1040 Schedule SE or had neither a W-2 nor 1040 Schedule SE on file. These individuals have a 58.4 percent predicted probability of reporting survey self-employment. Table 4 shows that individuals who receive the only W-2 issued by their firm's EIN have a 40.2 percent predicted likelihood of reporting survey self-employment. This predictive power declines sharply with firm size, though. Individuals who worked at a firm with two to three employees have a 24.2 percent predicted likelihood of reporting survey self-employment. Individuals who worked at a firm with four to nine employees have an 11.5 percent predicted likelihood of reporting survey self-employment. Thus, an administrative measure of firm size likely improves the accuracy of self-employment predictions for individuals who receive W-2 forms.

IV.C. Is Self-Employment Status Missing at Random?

The correlation between survey and administrative measures of self-employment status portrayed in Table 2 allows us to test the validity of the missing at random assumption that is necessary for hot deck imputation to yield unbiased estimates of self-employment status. This assumption requires that nonresponse to the self-employment status question is unrelated to self-employment status itself, conditional on the variables used to predict the missing self-employment status. We test this assumption in Table 3, which presents the results of a logit regression of survey nonresponse to self-employment status or incorporation status on administrative measures of self-employment status. This is a job-level regression run on our full sample, omitting only individuals who were not assigned a PIK.

In columns 1 through 3, we find no evidence that individuals who receive a 1040 Schedule SE are more or less likely to have missing survey self-employment status or incorporation status, relative to individuals who receive only W-2 forms and individuals who

have no administrative earnings records. This result holds regardless of whether we include no additional controls (column 1), we control for the variables used to predict self-employment status for hot deck imputation (column 2), or we control for additional variables (column 3).

But these findings change when we include indicators for all possible combinations of administrative earnings sources, including lack of administrative earnings. Columns 4, 5, and 6 of Table 3 show that individuals with no administrative earnings are much more likely to have missing self-employment status or incorporation status than individuals who receive only W-2 forms. After controlling for the variables used to predict self-employment status in hot deck imputation, individuals who have only 1040 Schedule SE forms are also more likely to have missing self-employment status or incorporation status than individuals who receive only W-2 forms. Given the correlations reported in Table 2, these findings suggest that the missing at random assumption is violated. Since SRMI controls for administrative earnings source when predicting missing self-employment status, it might restore the missing at random assumption.

IV.D. How Well Do the Imputation Methods Predict Aggregate Self-Employment?

Table 5 compares the distribution of self-employment status and incorporation status implied by each of the imputation methods to the reported distribution. All four methods match the reported distribution remarkably well. Perhaps the most important lesson from this table is that even though it does not include administrative data or an expanded set of predictor variables, the hot deck is sufficient for matching the reported distribution well.

IV.E. How Well Do the Imputation Methods Predict a Job's Reported Self-Employment Status?

Table 6 presents the percentage of jobs imputed their correct (*i.e.,* reported) self-employment and incorporation status. The reader should bear in mind that the model with the most correct predictions is not necessarily the best methodology. If accurate prediction was the goal imputation, then assigning all respondents salary employment would a better imputation technique than each technique examined. For example, assigning all jobs wage and salary employment would correctly predict 87.9 percent of the jobs' self-employment statuses compared to the 85.1 percent correct predicted by SRMI with administrative data. Instead the goal of imputation is to preserve the relationships between variables, so that an analysis of imputed data returns results identical to if it had been reported. Correctly predicted responses may be consistent with this goal but does not guarantee it is met.

The top row of Table 6 reveals that there is surprisingly little variation in the candidate methodologies' ability to match reported data across all jobs. The proportion imputed correctly ranges from 78.9 percent according to Bayesian Bootstrap without administrative data to 85.1 percent according to SRMI with administrative data. In judging how successfully these candidate methodologies impute, a useful benchmark is random assignment. If self-employment and imputation status were imputed randomly following the reported distribution, 68.3 percent of the imputed values would be correct across all jobs.

This seeming lack of variation across imputation methods for all jobs masks considerable variation across imputation methods for certain types of jobs. For example, 14.7 percent of self-employed jobs are imputed correctly according to Bayesian Bootstrap without administrative data. This is an improvement over random assignment, which only imputes 10.6 percent correctly. But Bayesian Bootstrap without administrative data is substantially worse than SRMI with administrative data, which imputes 48.5 percent of self-employed jobs correctly. No candidate imputation method predicts "other work arrangements" accurately, although SRMI with administrative data does perform slightly better than Bayesian Bootstrap without administrative data (5.5 percent imputed correctly compared with 1.9 percent).

We find evidence that the inclusion of administrative data improves accuracy more than the inclusion of more covariates. Both Bayesian Bootstrap with administrative data and SRMI without administrative data better predict self-employment statuses than Bayesian Bootstrap without administrative data. But Bayesian Bootstrap with administrative data outpaces SRMI without administrative data—44.5 percent correct versus 35.8 percent correct—and only slightly trails SRMI with administrative data—44.5 percent correct versus 48.5 percent correct.

In sum, while SRMI with administrative data appears to perform best for wage-employed and self-employed jobs, Bayesian Bootstrap with administrative data performs well too. The introduction of administrative data to the set of predictor variables appears to be more important than model-based imputation for improving imputation accuracy of some self-employment statuses.

IV.F. The Correlation of Imputed Values Across Methodologies

While Table 6 offers insight into what fraction of reported values each imputation method predicts correctly, it does not provide insight into the correlation of imputed values across imputation methods. For example, how much of the 14.7 percent of self-employed jobs

that Bayesian Bootstrap without administrative data imputes correctly does SRMI with administrative data also impute correctly? In other words, does Bayesian Bootstrap without administrative data impute a different set of self-employed jobs correctly than SRMI with administrative data does. Or are the jobs imputed correctly by Bayesian Bootstrap without administrative data merely a subset of the jobs imputed correctly by SRMI with administrative data? This question is especially important to consider when determining whether imputation accuracy declines systematically for any segment of the population when SRMI with administrative data replaces hot deck imputation.

To answer this question, Table 7 presents the fraction of self-employed jobs imputed according to each methodology that are also imputed to be self-employed jobs according to each other imputation methodology. Panel A performs this analysis on our full sample, while Panel B restricts attention to the set of jobs correctly imputed to self-employment. The first column of Panel A reveals a low degree of correlation between jobs imputed to self-employment according to Bayesian Bootstrap without administrative data and any other imputation method. No other imputation methodology imputes self-employment to more than 20 percent of the jobs imputed to self-employment by Bayesian Bootstrap without administrative data. The SRMI without administrative data model has the highest correlation with the SRMI with administrative model. SRMI without administrative data imputes self-employment to 50 percent of the jobs imputed to self-employment according to Bayesian Bootstrap with administrative data.

Panel B of Table 7 restricts attention to observations imputed correctly by a particular imputation method. Estimates along the diagonal demonstrate that there is substantial variability in imputed values across implicates. For example, among implicates for which Bayesian Bootstrap without administrative data correctly imputes self-employment, that same imputation method correctly imputes self-employment for a surprisingly low 14.7 percent of those individuals' other implicates. Other imputation methods are associated with higher correlation of imputed values across implicates, the highest being SRMI with administrative data. However, even among implicates for which SRMI with administrative data correctly imputes self-employment, that same imputation method correctly imputes self-employment for only about half of those individuals' other implicates.

Examining off-diagonal estimates in Panel B of Table 7 reveals that the jobs imputed correctly by Bayesian Bootstrap without administrative data are not merely a subset of the jobs

imputed correctly by SRMI with administrative data. Column 1 shows that among implicates for which Bayesian Bootstrap without administrative data correctly imputes self-employment, SRMI with administrative data correctly imputes only 7.5 percent of those implicates to self-employment. The within-implicate correlation between SRMI with administrative data and Bayesian Bootstrap without administrative data is thus 50.9 percent of the correlation across implicates of Bayesian Bootstrap without administrative data for these observations. By comparison, the within-implicate correlation between SRMI with administrative data and Bayesian Bootstrap without administrative data is 15.4 percent of the correlation across implicates of SRMI with administrative data for observations correctly imputed to self-employment by SRMI with administrative data.

Summing up, some information is likely lost when replacing hot deck imputation with SRMI with administrative data. However, this information loss is arguably outweighed by the increased accuracy of SRMI displayed in Table 6 and its lower cross-implicate variability displayed in Panel B of Table 7.

IV.G. Do the Imputation Methods Preserve Reported Correlations?

As a final analysis, we consider which correlations in reported data each self-employment imputation methodology preserves. To that end, we estimate a logit model of self-employment status on a large set of explanatory variables separately by gender. This regression is of special interest to self-employment researchers, and is commonly used to investigate the observable characteristics of self-employed individuals. The explanatory variables include race and ethnicity, marital status, partner's self-employment status, foreign born status, age of children, education, own age, disability status, net worth, region, metropolitan status, work schedule, an indicator for proxy response, and an indicator for being assigned a PIK.

Table 8 presents the predicted marginal effects of this regression for men.[12] The table presents five columns, each containing estimated marginal effects when the dependent variable comes from the following five sources: one for reported data and one for each of the candidate imputation methodologies.[13] Regressions of imputed self-employment status on observables use

[12] We have also estimated the model for women. Estimates are available in Appendix Table 1. The same general result holds: SRMI with administrative data preserves some correlations found in reported data that Bayesian Bootstrap without administrative data does not, while some correlations found in reported data are not captured by any candidate imputation method.

[13] We have also estimated a multinomial logit model where the dependent variable distinguishes between incorporated and unincorporated self-employed individuals. Estimates are available upon request.

only the first implicate of self-employment status. So unlike those in Tables 5 through 7, standard errors in columns 2 through 5 of Table 8 do not reflect the uncertainty introduced by imputation.

Table 8 reveals that for men, incorporating model-based imputation and including administrative records in the set of predictor variables can preserve some correlations that hot deck imputation does not preserve. First, according to the Bayesian Bootstrap without administrative data there is positive correlation between marriage and self-employment, but in the reported data and models with administrative data a negative correlation is observed. Even though Bayesian Bootstrap without administrative data conditions on marital status when matching donors, the data imputed according to this method reflects the correlation between marital status and self-employment status relatively poorly. Perhaps because it does not capture the correlation between having a self-employed partner and self-employment status meaning it does not distinguish an important underlying component of the correlation between marriage and self-employment.

Second, Bayesian Bootstrap without administrative data also reflects the correlation between foreign-born status and self-employment status relatively poorly. Immigrants are less likely than native-born individuals to be self-employed in data imputed according to Bayesian Bootstrap without administrative data, while immigrants are more likely than native-born individuals to be self-employed in reported data. The inability of Bayesian Bootstrap without administrative data to preserve this relationship is not surprising because it does not use foreign-born status to match donors to nonrespondents. By contrast, self-employment status imputed according to the SRMI model with administrative data preserves the positive correlation between foreign-born status and self-employment status, though we reject the hypothesis that the marginal effects SRMI model and the reported model are equal.

Third, Bayesian Bootstrap without administrative data reflects the correlation between self-employment and work schedule relatively poorly. We reject the hypothesis of equality between the coefficients estimated on reported self-employment status and self-employment status imputed according to Bayesian Bootstrap without administrative data for several work schedule groups: those performing evening shift work, those working Monday through Friday, those working Saturday and Sunday, those working more than 50 hours in a usual week, and those who sometimes worked from home. By contrast, we only reject the hypothesis of equality

between the coefficients estimated on reported self-employment status and self-employment status imputed according to SRMI with administrative data for fewer work schedule groups: those working Monday through Friday, those working more than 50 hours in a usual week, and those sometimes working from home. But it is worth noting that the SRMI with administrative data model does capture the positive correlations for each of these groups. The ability of SRMI with administrative data to reproduce more correlations between self-employment status and work schedule present in reported data is especially surprising because work schedule does not enter the model directly.

While Table 8 reveals some promise of administrative data and model-based imputation to preserve more correlations present in reported data, it also reveals evidence that all four candidate imputation methods reflect poorly some correlations present in reported data. For example, we reject the hypothesis that coefficient estimates on indicators for married and cohabiting are equal between reported data and imputed data for all four candidate imputation methods. Similarly, we reject the hypothesis that the coefficient estimate on age is equal between reported data and imputed data for all four candidate imputation methods. Table 8 also shows negative association between being Black and self-employment in all imputation models, but no statistically significant relationship in the reported data.

Perhaps the most striking correlation in the reported data is the extraordinarily predictive power of having a self-employed spouse. The SRMI with administrative data does preserve this relationship between spousal and own self-employment status, but its marginal effect misses the reported estimate by a wide margin. But the Bayesian Bootstrap without administrative data model fails to capture any statistically significant relationship between spousal and own self-employment status.

V Conclusion

Historically, missing self-employment data have been imputed by hot deck in U.S. Census Bureau surveys. A methodology that inherently limits the set of variables used to match nonrespondents with donors to relatively few variables, typically demographic in nature. Imputed self-employment data thus presents an understudied threat to identification for self-employment researchers. Coefficient estimates on any variables not used in this matching algorithm are likely to be biased towards zero in self-employment status or self-employment transition regressions. And estimates of self-employment rates will suffer from nonresponse bias

if self-employment nonresponse is related to self-employment status itself, conditional on the variables used to match nonrespondents to donors.

In this paper, we describe a new imputation methodology which aims to mitigate these empirical challenges. Sequential Regression Multiple Imputation——or its relative Sequential Regression Multivariate Imputation—enables data producers to predict missing self-employment status and incorporation status via regression model with an inclusive set of explanatory variables. These explanatory variables may include administrative records of self-employment status that are available for respondents and nonrespondents alike.

We show that although the correlation between survey and administrative measures of self-employment status are imperfect, administrative measures of self-employment contain important information that can improve predictions. Consequently, although hot deck imputation predicts self-employment status correctly more often than would be expected by random assignment, imputation models that leverage administrative measures of self-employment predict self-employment status correctly much more often than hot deck imputation does.

We also demonstrate that when regressing self-employment status on a host of observables, SRMI with administrative data preserves more of the correlations present in reported data relative to traditional hot deck imputation. Nevertheless, the feasible set of predictor variables is too large to capture every possible relationship with self-employment status and we find evidence of match bias.

Therefore, while we conclude the new imputation technique implemented in the 2018 Survey of Income and Program Participation constitutes an improvement over the hot deck, we recommend that analysts make careful decisions about how to handle respondents with imputed self-employment statuses. We recommend that analysts estimate their models once with imputed self-employment and incorporation status responses and again without imputed responses, acknowledging and interpreting any meaningful differences in coefficient estimates. We also recommend estimating a model on respondents only, re-weighting to account for response probability. Administrative data should enter this re-weighting algorithm if available. While none of these approaches would likely fully address match and nonresponse bias, the combination of approaches would allow researchers to assess their likely impact.

References

Abraham, Katherine G., John C. Haltiwanger, Kristin Sandusky, and James R. Spletzer (2021), "Measuring the Gig Economy: Current Knowledge and Open Issues". In *Measuring and Accounting for Innovation in the 21st Century*. Ed. Carol Corrado, Jonathan Haskel, Javier Miranda, and Daniel Sichel: University of Chicago Press.

Abraham, K. G., Haltiwanger, J. C., Hou, C., Sandusky, K., & Spletzer, J. R. (2021). "Reconciling Survey and Administrative Measures of Self-Employment." *Journal of Labor Economics*, 39(4), 825-860.

Andridge, Rebecca R. and Little, Roderick, J. A. (2010), "A Review of Hot Deck Imputation for Survey Non-Response" International Statistical Review 78 (1).

Benedetto, Gary, Joanna Motro, and Martha Stinson (2015), "Introducing Parametric Models and Administrative Records into 2014 SIPP Imputations." Proceedings of the 2015 Federal Committee on Statistical Methodology Research Conference.

Bollinger, Christopher R. and Barry T. Hirsch (2006) "Match Bias from Earnings Imputation in the Current Population Survey: The Case of Imperfect Matching." *Journal of Labor Economics* 24(3), 483–519.

Bollinger, Christopher R., Barry T. Hirsch, Charles M. Hokayem, and James P. Ziliak (2019) "Trouble in the Tails? What We Know about Earnings Nonresponse 30 Years after Lillard, Smith, and Welch" *Journal of Political Economy* 127(5) 2143–2185.

Cantwell, P. J., Hogan, H., & Styles, K. M. (2005). Imputation, Apportionment, and Statistical Methods in the U.S. Census: Issues Surrounding Utah v. Evans. U.S. Census Bureau.

Eggleston, Jonathan, and Lori Reeder (2018), "Does Encouraging Record Use for Financial Assets Improve Data Accuracy? Evidence from Administrative Data." *Public Opinion Quarterly* 82(4), 686–706.

Giefer, Katherine, Abby Williams, Gary Benedetto, and Joanna Motro (2015), "Program Confusion in the 2014 SIPP: Using Administrative Records to Correct False Positive SSI Reports." Proceedings of the 2015 Federal Committee on Statistical Methodology Research Conference.

Hirsch, Barry T. and Edward J. Schumacher (2004), "Match Bias in Wage Gap Estimates Due to Earnings Imputation." *Journal of Labor Economics* 22(3), 689–722.

Light, Audrey and Robert Munk (2018), "Business Ownership versus Self-Employment." *Industrial Relations* 57(3), 435–468.

Meyer, Bruce D., Wallace K. C. Mok, and James X. Sullivan (2015), "Household Surveys in Crisis." *Journal of Economic Perspectives* 29(4), 199–226.

Ono, M., & Miller, H. P. (1969). Income Nonresponses in the Current Population Survey. *ASA Proc Social Statistics*, 277-299
Raghunathan, Trivellore E., James M. Lepkowski, John Van Hoewyk, and Peter Solenberger (2001), "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models." *Survey Methodology* 27(1), 85–95.

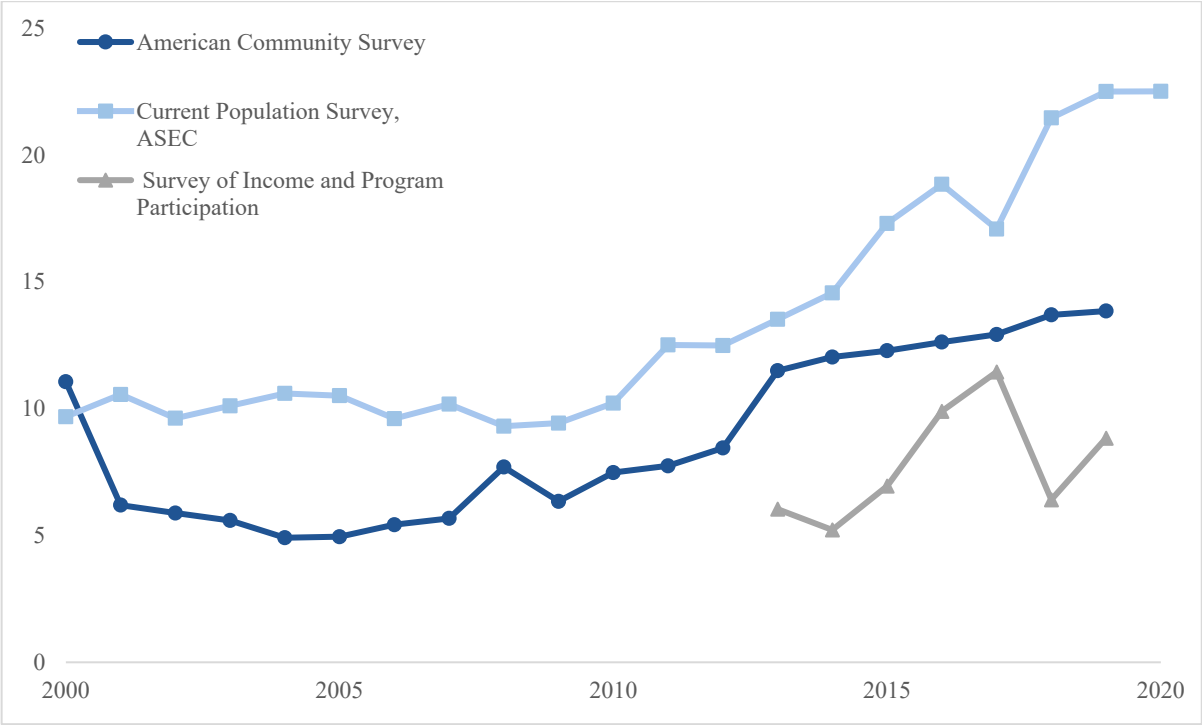
Sage Publications. (2008). Hot Deck Imputation. In P. J. Lavrakas, *Encycloedia of Survey Research Methods* (pp. 315-317).

Slemrod, Joel (2007) "Cheating Ourselves: The Economics of Tax Evasion" *Journal of Economic Perspectives* 21(1) 25-48.

Theis, Thomas N. and HSP Wong (2017) "The End of Moore's Law: A New Beginning for Information Technology" Computing in Science and Engineering 19 (2) 41-50.

Wagner, Deborah, and Mary Layne (2014) "The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications (CARRA) Record Linkage Software." CARRA Working Paper #2014

Figure 1: Percent of In Universe Observations with Imputed Self-Employment Status by Census Household Survey



Sources: American Community Survey 2000-2019: Steven Ruggles, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler and Matthew Sobek. IPUMS USA: Version 11.0 [dataset]. Minneapolis, MN: IPUMS, 2021. https://doi.org/10.18128/D010.V11.0

Annual Social and Economic Supplement, Current Population Survey – ASEC 2000-2020: Sarah Flood, Miriam King, Renae Rodgers, Steven Ruggles, J. Robert Warren and Michael Westberry. Integrated Public Use Microdata Series, Current Population Survey: Version 9.0 [dataset]. Minneapolis, MN: IPUMS, 2021. https://doi.org/10.18128/D030.V9.0

Survey of Income and Program Participation Reference Years 2013-2019

Note: The data is unweighted to show the percent of allocated responses on the public use microdata by year. The CPS-ASEC imputation rate includes whole supplement imputation.

Table 1: The Impact of Sample Selection Criteria on Sample Size

Selection Criteria	Sample Size
Full sample of jobs	40,000
Drop if entire job imputed	38,500
Drop if self-employment status imputed	38,000
Drop if incorporation status imputed	38,000
Drop if military industry or occupation	37,500
Drop if lives outside 50 states or D.C.	37,500

Source: U.S. Census Bureau, Survey of Income and Program Participation, 2014 Panel, Wave 1

Note: All counts are rounded according to the guidelines of the U.S. Census Bureau Disclosure Review Board and estimates are rounded to four significant digits. Our population of interest is identical to our full analysis sample: SIPP respondents with reported self-employment and incorporation status, who were not in a military industry or occupation, and who did not live outside the 50 states and the District of Columbia. Consequently, inferences drawn in this paper are descriptive of the SIPP 2014 respondents and not necessarily representative of the U.S. population at large. All estimates in this paper are unweighted and do not account for the sample design.

Table 2: Cross-Tabulation of Self-Employment Status in the Survey of Income and Program Participation (SIPP) and the Detailed Earnings Record (DER) at the Person Level

	Tax Forms Observed in the DER				Number of SIPP
In the SIPP, the respondent works for/is…	Both W-2 and 1040-SE	Only W-2	Only 1040-SE	Neither	Respondents
Employer only	1,000	24,000	350	800	26,500
Row percent	3.77	90.57	1.32	3.02	
Column percent	62.50	94.12	21.88	44.44	86.89
Employer and some other arrangement	*Suppressed*				
Row percent					
Column percent					
Unincorporated self-employed only	150	200	950	600	1,900
Row percent	7.89	10.53	50.00	31.58	
Column percent	9.38	0.78	59.38	33.33	6.23
Unincorporated self-employed with more than one job†	250	300	40	30	600
Row percent	41.67	50.00	6.67	5.00	
Column percent	15.63	1.18	2.50	1.67	1.97
Incorporated self-employed only	90	500	200	200	1,000
Row percent	9.00	50.00	20.00	20.00	
Column percent	5.63	1.96	12.50	11.11	3.28
Incorporated self-employed with more than one job	*Suppressed*				
Row percent					
Column percent					
Some other arrangement only	30	100	60	150	350
Row percent	8.57	28.57	17.14	42.86	
Column percent	1.88	0.39	3.75	8.33	1.15
Column total	1,600	25,500	1,600	1,800	30,500
Row percent	5.25	83.61	5.25	5.90	

Source: U.S. Census Bureau, Survey of Income and Program Participation, 2014 Panel, Wave 1 and Social Security Administration, Detailed Earnings Record 2013

Note: Source 2014 SIPP wave 1. All totals are rounded to satisfy Census DRB rules, implying percentages may not sum to 100. Estimates in this table are person-level, so sample size will differ from the sample size cited in Table 1 which includes multiple jobs per person. Our population of interest is identical to our full analysis sample: SIPP respondents with reported self-employment and incorporation status, who were not in a military industry or occupation, and who did not live outside the 50 states and the District of Columbia. Consequently, inferences drawn in this paper are descriptive of the SIPP 2014 respondents and not necessarily representative of the U.S. population at large. All estimates in this paper are unweighted and do not account for the sample design.

†Persons with both incorporated and unincorporated self-employed jobs are placed in the incorporated self-employed with more than one job category. However, this is quite rare and does not meaningfully change the row and column percentages

Table 3: Do Tax Forms Predict Missing Self-Employment Status or Incorporation Status? Logit Marginal Effects of Tax Forms on the Probability of Self-Employment or Incorporation Status Imputation

Filed a 1040-SE	-0.522	0.129	0.0869			
	(0.336)	(0.378)	(0.438)			
Filed both 1040-SE and W2				-0.587	-0.144	0.0400
				(0.414)	(0.451)	(0.477)
Filed only a W2				---	---	---
				---	---	---
Filed only a 1040-SE				0.474	1.513**	2.235***
				(0.520)	(0.618)	(0.811)
Filed neither a W2 or a 1040-SE				6.292***	6.458***	6.622***
				(0.692)	(0.711)	(0.766)
N	37,000	37,000	37,000	37,000	37,000	37,000
No Controls	Yes	No	No	Yes	No	No
Hot Deck Controls	No	Yes	No	No	Yes	No
Hot Deck Plus Controls	No	No	Yes	No	No	Yes

Source: U.S. Census Bureau, Survey of Income and Program Participation, 2014 Panel, Wave 1

Note: The sample for this regression is our full sample, omitting only individuals who were not assigned a Protected Identification Key (PIK). All counts are rounded according to the guidelines of the U.S. Census Bureau Disclosure Review Board and estimates are rounded to four significant digits. Standard errors in parentheses.

*** p<0.01, ** p<0.05, * p<0.1

Table 4: The Number of Employees at an EIN Predicts Self-Employment

Number of employees at EIN	Predicted percent self-employed from Linear Probability Model
Missing	58.39***
	(0.851)
1	40.16***
	(3.063)
2 to 3	24.22***
	(1.825)
4 to 9	11.54***
	(0.884)
10 to 25	4.815***
	(0.515)
26 to 49	2.498***
	(0.461)
50 to 100	2.170***
	(0.399)
101 to 200	0.899***
	(0.337)
201 to 500	0.741***
	(0.264)
500 to 1,000	1.202***
	(0.316)
Over 1,000	2.082***
	(0.144)
N	35,000

Source: U.S. Census Bureau, Survey of Income and Program Participation, 2014 Panel, Wave 1

Note: All counts are rounded according to the guidelines of the U.S. Census Bureau Disclosure Review Board and estimates are rounded to four significant digits. The sample excludes nonrespondents, respondents with a military industry or occupation, respondents living outside the 50 states and DC, and respondents not assigned a Protected Identification Key (PIK), who are not linkable to W-2 tax records. A missing firm size indicates individuals who were linkable to W-2 tax records yet received no W-2s in 2013. Predicted percentages are generated from a linear probability model regressing a job's self-employment status on person-level EIN indicators. Inferences drawn in this paper are descriptive of the SIPP 2014 respondents and not necessarily representative of the U.S. population at large. All estimates in this paper are unweighted and do not account for the sample design. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 5: The Percent of Jobs in Each Self-Employment Status by Reporting Status and Imputation Model

Data is/imputed using	Reported	Bootstrap	Bootstrap	SRMI	SRMI
Administrative data used in imputation…	n/a	No	Yes	No	Yes
Wage-employed	87.85	88.13	87.64	87.75	87.57
		(0.26)	(0.29)	(0.24)	(0.28)
Self-employed, All	10.58	10.51	11.17	10.55	10.78
		(0.24)	(0.28)	(0.18)	(0.23)
Self-employed, Incorporated	3.36	3.54	3.85	3.30	3.64
		(0.16)	(0.21)	(0.20)	(0.18)
Self-employed, Unincorporated	7.22	6.96	7.33	7.24	7.14
		(0.19)	(0.21)	(0.19)	(0.18)
Some other arrangement	1.57	1.36	1.18	1.70	1.64
		(0.12)	(0.19)	(0.12)	(0.11)

Source: U.S. Census Bureau, Survey of Income and Program Participation, 2014 Panel, Wave 1

Note: Estimates are averaged across five implicates of missing data and are rounded to four signifcant digits. Our population of interest is identical to our full analysis sample: SIPP respondents with reported self-employment and incorporation status, who were not in a military industry or occupation, and who did not live outside the 50 states and the District of Columbia. Consequently, inferences drawn in this paper are descriptive of the SIPP 2014 respondents and not necessarily representative of the U.S. population at large. All estimates in this paper are unweighted and do not account for the sample design. Standard errors in parentheses are based on random variation across implicates.

Table 6: The Percent of Jobs with the Correct Self-Employment Status Imputed by Imputation Model

Data imputed using…	Bootstrap	Bootstrap	SRMI	SRMI	Random
Administrative data used in imputation…	No	Yes	No	Yes	Assignment*
All jobs	78.87	84.15	82.59	85.13	68.27
	(0.34)	(0.32)	(0.21)	(0.32)	
Wage-employed jobs	88.71	92.25	91.10	92.86	87.85
	(0.31)	(0.27)	(0.19)	(0.25)	
Self-employed jobs	14.69	44.53	35.76	48.48	10.58
	(1.06)	(1.61)	(1.18)	(1.70)	
Self-employed, incorporated jobs	6.33	9.92	10.40	18.15	3.36
	(1.17)	(1.46)	(1.12)	(3.07)	
Self-employed, unincorporated jobs	9.58	37.91	29.92	39.61	7.22
	(0.90)	(1.41)	(1.08)	(1.12)	
Some other arrangement jobs	1.93	2.53	3.24	5.47	1.57
	(0.69)	(1.01)	(1.16)	(1.36)	

Source: U.S. Census Bureau, Survey of Income and Program Participation, 2014 Panel, Wave 1

Note: Estimates are averaged across five implicates of missing data and are rounded to four significant digits. Our population of interest is identical to our full analysis sample: SIPP respondents with reported self-employment and incorporation status, who were not in a military industry or occupation, and who did not live outside the 50 states and the District of Columbia. Consequently, inferences drawn in this paper are descriptive of the SIPP 2014 respondents and not necessarily representative of the U.S. population at large. All estimates in this paper are unweighted and do not account for the sample design. Standard errors in parentheses are based on random variation across implicates.

*The expected percentage of jobs with a correctly assigned class of worker status based on complete random assignment. The expected values are calculated using the values in the reported column from Table 4.

Table 7: The Correspondence of the Imputation Methodologies at the Job Level

Panel A: Percent of Jobs Imputed to Be Self-Employed, Conditional Upon Another Methodology Imputing Self-Employment

	Bootstrap no admin	Bootstrap admin	SRMI no admin	SRMI admin
Column imputed job as self-employed…	Yes	Yes	Yes	Yes
Column imputed job required to match reported status…	No	No	No	No
Bootstrap, no administrative data	10.51	1.52	1.47	1.51
	(0.24)	(0.07)	(0.11)	(0.10)
Bootstrap, administrative data	1.52	11.17	2.76	4.44
	(0.07)	(0.28)	(0.13)	(0.19)
SRMI, no administrative data	1.47	2.76	10.55	5.24
	(0.11)	(0.13)	(0.18)	(0.13)
SRMI, administrative data	1.51	4.44	5.24	10.78
	(0.10)	(0.19)	(0.13)	(0.23)

Panel B: Percent of Jobs Imputed to Be Self-Employed, Conditional Upon Another Methodology Imputing Self-Employment and That Imputation Matching the Reported Value

	Bootstrap no admin	Bootstrap admin	SRMI no admin	SRMI admin
Column imputed job as self-employed…	Yes	Yes	Yes	Yes
Column imputed job required to match reported status…	Yes	Yes	Yes	Yes
Bootstrap, no administrative data	14.69	6.76	5.60	7.47
	(1.06)	(0.57)	(0.69)	(0.61)
Bootstrap, administrative data	6.76	44.53	17.99	28.37
	(0.57)	(1.61)	(1.26)	(1.54)
SRMI, no administrative data	5.60	17.99	35.76	25.10
	(0.69)	(1.26)	(1.18)	(1.02)
SRMI, administrative data	7.47	28.37	25.10	48.48
	(0.61)	(1.54)	(1.02)	(1.70)

Source: U.S. Census Bureau, Survey of Income and Program Participation, 2014 Panel, Wave 1

Note: Estimates are averaged across five implicates of missing data and are rounded to four signifcant digits. Our population of interest is identical to our full analysis sample: SIPP respondents with reported self-employment and incorporation status, who were not in a military industry or occupation, and who did not live outside the 50 states and the District of Columbia. Consequently, inferences drawn in this paper are descriptive of the SIPP 2014 respondents and not necessarily representative of the U.S. population at large. All estimates in this paper are unweighted and do not account for the sample design. Standard errors in parentheses are based on random variation across implicates.

Table 8: Logit Marginal Effects of Variables on the Probability of Self-Employment for Men, Marginal Effects are Multiplied by 100

Data is/imputed using		Reported	Bootstrap	Bootstrap	SRMI	SRMI
Administrative data used in imputation…		n/a	No	Yes	No	Yes
Race/Ethnicity	Black	-0.0553	-4.129***	-3.341***	-2.132***	-2.488***
		(0.476)	(0.717)	(0.743)	(0.768)	(0.747)
	White, Hispanic	-0.0670	-1.227	-2.023**	-3.302***	-2.852***
		(0.497)	(0.857)	(0.808)	(0.802)	(0.762)
	Some other race	-1.266**	-4.606***	-4.049***	-1.685*	-2.237**
		(0.564)	(0.885)	(0.878)	(0.975)	(0.933)
Nativity	Foreign Born	1.531***	-1.779**	4.625***	2.146**	4.110***
		(0.482)	(0.807)	(0.946)	(0.890)	(0.890)
Marital Status	Cohabitating	-7.463***	2.995**	-2.236**	-2.240**	-3.719***
		(0.514)	(1.206)	(0.908)	(0.960)	(0.849)
	Married	-9.412***	3.068***	-3.009***	-0.608	-4.700***
		(0.470)	(0.794)	(0.808)	(0.810)	(0.785)
	Divorced	0.678	-0.396	0.918	0.936	1.441
		(0.436)	(1.054)	(1.030)	(1.020)	(0.954)
Partner	Partner self-employed	73.84***	-0.343	21.71***	14.45***	27.06***
		(0.990)	(0.741)	(1.222)	(1.045)	(1.240)
Children	Has a child under 6	0.00378	-1.695**	1.319*	-0.533	-0.990
		(0.475)	(0.688)	(0.740)	(0.723)	(0.693)
	Has a child under 18	0.183	0.0907	0.342	-0.587	0.325
		(0.355)	(0.565)	(0.557)	(0.551)	(0.540)
Education	Less than high school	1.375***	1.723*	4.856***	1.759*	2.551***
		(0.526)	(0.982)	(1.005)	(0.928)	(0.878)
	Some College, no degree	-0.328	0.512	1.952***	-1.414**	-1.975***
		(0.400)	(0.728)	(0.755)	(0.673)	(0.625)
	Two-Year degree	-0.369	2.066**	0.703	-0.931	-2.212***
		(0.551)	(0.993)	(0.958)	(0.888)	(0.832)
	Bachelors degree	-1.218***	0.892	0.130	-0.243	-1.790***
		(0.407)	(0.735)	(0.739)	(0.689)	(0.639)

	Advanced degree	-2.371***	1.291	0.456	-0.0699	-0.427
		(0.471)	(0.851)	(0.880)	(0.796)	(0.768)
Age	Age/10	1.583***	3.103***	2.076***	4.061***	3.208***
		(0.122)	(0.222)	(0.231)	(0.211)	(0.199)
Health	Any disability	0.0915	0.301	0.540	0.831	0.270
		(0.460)	(0.754)	(0.757)	(0.730)	(0.704)
Wealth	Inverse hyperbolic sine of respondent net worth	1.629***	0.0470	0.319	-0.634*	0.741**
		(0.237)	(0.343)	(0.343)	(0.340)	(0.353)
Geography	Living in a metropolitan area	-0.336	-0.259	0.0833	-0.340	-0.360
		(0.360)	(0.574)	(0.576)	(0.573)	(0.556)
	Midwest	-0.880*	-0.302	-0.828	-2.013***	-1.715**
		(0.486)	(0.800)	(0.800)	(0.753)	(0.738)
	South	0.348	-0.361	0.126	-1.174	0.251
		(0.447)	(0.749)	(0.757)	(0.725)	(0.710)
	West	0.474	0.727	-0.00880	-0.785	-0.206
		(0.498)	(0.866)	(0.854)	(0.804)	(0.789)
Schedule	Evening shift	-2.900***	0.361	-1.954**	-0.462	-2.976***
		(0.555)	(0.868)	(0.832)	(0.871)	(0.797)
	Other shift	0.887**	0.127	0.0869	1.223*	0.768
		(0.375)	(0.657)	(0.632)	(0.653)	(0.596)
	Worked Monday-Friday	2.911***	-0.198	1.458**	1.237**	1.431**
		(0.411)	(0.655)	(0.632)	(0.626)	(0.602)
	Worked Saturday-Sunday	2.596***	0.289	1.275**	0.876	1.979***
		(0.400)	(0.653)	(0.637)	(0.635)	(0.612)
	Worked 20 or fewer hours	2.103***	-0.00296	0.845	0.537	0.574
		(0.596)	(0.945)	(0.932)	(0.898)	(0.848)
	Worked 35 or fewer hours	4.863***	4.211**	2.043	4.344**	3.714*
		(1.743)	(2.141)	(1.946)	(2.084)	(1.957)
	Worked 40 hours or more	0.848	2.258	-1.058	-0.361	-1.145
		(1.328)	(1.655)	(1.805)	(1.759)	(1.732)
	Worked 50 hours or more	3.657***	0.334	1.720***	0.426	2.181***
		(0.426)	(0.626)	(0.648)	(0.624)	(0.613)

	Sometimes worked from home	7.846***	-1.494*	6.533***	2.543***	4.059***
		(0.792)	(0.868)	(1.041)	(0.915)	(0.896)
	Primarily worked from home	1.179	2.765*	-0.736	-1.052	0.785
		(0.744)	(1.535)	(1.087)	(1.092)	(1.109)
Survey Variables	Is Respondent PIKed	-0.0288	0.410	2.378**	-1.495	-0.0208
		(0.544)	(1.051)	(0.936)	(1.074)	(0.985)
	Proxy response	0.0518	-0.572	-0.275	0.943*	0.711
		(0.346)	(0.527)	(0.539)	(0.534)	(0.524)
N		19000	19000	19000	19000	19000

Source: U.S. Census Bureau, Survey of Income and Program Participation, 2014 Panel, Wave 1
Note: All counts are rounded according to the guidelines of the U.S. Census Bureau Disclosure Review Board and estimates are rounded to four significant digits. Our population of interest is identical to men in our full analysis sample: male SIPP respondents with reported self-employment and incorporation status, who were not in a military industry or occupation, and who did not live outside the 50 states and the District of Columbia. Consequently, inferences drawn in this paper are descriptive of the SIPP 2014 respondents and not necessarily representative of the U.S. population at large. All estimates in this paper are unweighted and do not account for the sample design. Regressions of imputed self-employment status on observables use only the first implicate of self-employment status, so standard errors do not reflect the uncertainty introduced by imputation.

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1