Imputations: Benefits, Risks and a Method for Missing Data

Nikolas Mittag*

Harris School Of Public Policy, University of Chicago

May 17, 2013

Abstract: Missing data is a frequent problem in economics, either because some variables are missing from a data set or values are missing for some observations. Researchers usually either omit the affected variables and observations or impute them. While the consequences of the former are well understood, the imputation and missing data literature has focused on the conditions under which they lead to unbiased estimates. These conditions often do not hold, but there is little evidence on the circumstances under which missing data methods improve estimates if the conditions for unbiased estimates are violated. I first examine these conditions by discussing the circumstances under which missing data methods can be beneficial and frequent sources of bias. I then discuss advantages and problems of common missing data methods. Two important problems are that most methods work well for some models, but poorly for others and that researchers often do not have enough information to use imputed observations in common data sets appropriately. To address these problems, I develop a method based on the conditional density that works well for a wide range of models and allows producers of the data to incorporate private information and expertise, but still allows users of the data to adjust the imputations to their application and use them appropriately. Applications to common problems show that the conditional density method works well in practice.

Keywords: missing data; imputation; data combination.

JEL Classification Numbers: C18, C80, C83

*Address: Harris School of Public Policy, University of Chicago, 1155 E. 60th Street, Chicago, IL 60637, mittag@uchicago.edu

1 Introduction

Missing data can occur in the form of missing values and missing variables and both are frequent problems in economics. Missing values occur when a variable is partly missing because the required information could not be obtained for some units. The most common cause for missing values in surveys is non-response, which is prevalent in any survey and can be severe. Non-response can be refusal to answer the survey at all (unit non-response) or refusal to answer specific questions (item non-response). The rates for both types of non-response vary greatly by survey. For example, the Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC) has low rates of unit non-response (about 8%), but item non-response exceeds 50% even on some components of the core income questions. Entire variables may be missing, for example, because the relevant questions were not included in a survey or the information is confidential. Surveys are usually designed for a specific purpose and often do not contain variables that are not required for this purpose. For example, the CPS collects a lot of information about income, but does not include detailed questions on assets. These variables are important in applications such as the estimation of take-up models of government programs, but are often omitted because they are not available in the data.

A common way to proceed in the presence of missing data is to analyze the complete data only. If entire variables are missing from the data, this amounts to omitting the variables from the model. In the case of missing values, the analysis is usually performed on complete cases, i.e. units for which all relevant variables are available. Another common way to deal with missing data is to create a "complete" data set by allocating values to the missing observations. Some methods assign multiple values per observation or avoid assigning explicit values altogether, but even though it usually refers to methods that assign explicit values, in lieu of a better term, I use "imputation" to refer to all methods that attempt to solve the problem of missing data. Common economic surveys already include imputed values for the units that did not respond. These values are usually predicted based on the respondents, which is not feasible if the variables are missing entirely. If the missing variables are available in another dataset, however, the entire variables can be imputed. The main difference to partly missing variables is that the source data for the imputations is from a different data set or time period in the latter case.

Authors of articles in leading economic journals are roughly evenly split between including and excluding units with imputed values from their analysis. Besides showing that both approaches are frequent, the fact that different studies on similar issues using the same data set often differ in their choice to use imputed values suggests that this decision is usually subjective. The choice to include or exclude imputed values is rarely justified by explicit arguments and often not even mentioned, which supports the notion that it is often made arbitrarily. This paper attempts to provide information on the potential risks and benefits of

imputations in order to enable researchers to make more informed decisions when deciding to omit missing variables and observations or to use an imputation method. For most imputation methods, the conditions under which estimation will be unbiased are known, but rarely plausible in practice. If these conditions are violated, the researcher usually faces a choice between bias from omitting missing data or bias from imputation. Consequently, the choice is between two biased estimators and the relevant question is which bias is more severe, i.e. whether including the imputations will improve the estimates.

While the consequences of omitting variables or observations from a model are well understood, less is known about the advantages and disadvantages of imputation methods when their assumptions are violated. In order to enable more informed judgments, I first discuss potential benefits from missing data methods and the conditions under which they can occur in section 2. Since missing data methods usually make additional assumptions that can lead to bias when violated, they should only be used if there are potential benefits. These benefits have to be weighted against the potential bias, so I line out common sources of bias and how they can be avoided in section 3. This paper focuses on the advantages and problems of imputations when the information used to impute a variable that is (partially) missing from the data is obtained from complete observations or an additional data set. However, the problems and advantages are expected to extend to other settings in which imputation methods are commonly applied, such as predicting factor or IRT scores or data augmentation to avoid selection or measurement error (e.g. Brownstone and Valletta, 1996).

Whether the benefits are likely to outweigh the costs of imputations depends on the imputation method used and has implications for the ideal imputation method in specific cases, so section 4 evaluates how well common missing data methods capture the benefits and avoid the problems. Two important conclusions from these analyses are that researchers need more information on the way imputations are done in order to use them appropriately and that most common imputation methods are ideal for some models, but fare poorly in others. In section 5 I discuss a method based on an estimate of the conditional density of the missing variable that has several desirable properties. Most importantly, it works well for a wide range of models and allows both the provider and the user of the data set to incorporate private information in the imputation process. This is important if the data provider wants to make imputations available. I show that the method performs well in practice in section 5.2.

2 When are Imputations desirable?

Prior to the question on how imputations should be done and used, the researcher needs to decide whether imputations should be used at all. This section examines under which conditions imputations could be beneficial and under which they cannot. Since imputations require additional assumptions and can cause

bias, they should not be used in the latter case. If there are potential benefits, they need to be compared to the potential biases, which are discussed in section 3. This section abstracts from the imputation method in that it assumes that is is ideal for the model at hand. After introducing the notation, I show that if the variable is entirely missing, imputations are particularly appealing if the source data is comparable to the outcome data. If the variable is only partially missing, benefits mainly arise if missing values are non-ignorable and the imputations contain information from outside the model, in other cases imputations should usually be avoided. I show that these conditions have implications for desirable features of imputation methods.

Throughout this paper, I assume that there are two sets of variables: variables used in the outcome model, X, and variables used in the imputation model, Z. The "conditioning variables" Z may contain a subset of the "model" variables X. To keep the discussion general, X may include dependent and independent variables, but when discussing specific models that contain both types of variables, X refers to the independent variables and Y to the dependent variable(s). The researcher wants to estimate some statistic $Q(X)$, but one of the model variables is partially or entirely missing: $X = (X_{Obs}, X_{Mis})$ and the researcher only observes a sample from X_{Obs}, which may be empty if the variable is missing entirely. The extension to the empirically important case where different variables are missing for different observations is conceptually straight forward, but notationally cumbersome, so it is not formalized here. If all variables were observed, the researcher would estimate $Q()$ by $\hat{Q}(X)$. The researcher can use the random variable X_{Imp} instead of the unobserved X_{Mis}, thereby working with $\tilde{X} = (X_{Obs}, X_{Imp})$ instead of X. Usually, X_{Imp} are the predicted values of X_{Mis} based on Z_{Mis} and an estimated model of the relationship between X_{Mis} and Z_{Mis}. The data from which the imputation model is obtained is the "source data", while the data that is used to estimate $Q()$ is the "outcome data". The source data can be a subset of the outcome data, as is common with imputations for non-response, but it can also be a separate dataset or even multiple data sets or known parameters. Let $\eta = \tilde{X} - X$ denote the "imputation error", which is 0 for X_{Obs} and the variables in X that do not contain missing values. If X_{Obs} is not empty, an alternative is to estimate Q by $\hat{Q}(X_{Obs})$, i.e. to use only complete cases.

2.1 Imputing a Missing Variable

When a variable is missing from the data entirely, the alternative to imputation is to omit the variable at stake from the model entirely. This is only an option if a control variable is imputed and the coefficient on the variable is not of interest. If there is no bias from omitting the variable, the only reason to impute it would be on grounds of efficiency. An entire variable is rarely imputed for reasons of efficiency alone, so I

focus on situations in which bias is the main concern. If there is omitted variable bias, imputing the missing variable is an improvement as long as the bias from imputation is smaller than the omitted variable bias. I discuss conditions under which there is no bias from imputation, but if these conditions do not hold, the trade-off is similar to that of including a badly measured proxy instead of a true variable and depends on how severely the assumptions are violated.

Several papers have proposed methods to impute a variable from another data set (see Ridder and Moffitt, 2007, for a review) and have discussed conditions under which these methods yield unbiased estimates. A sufficient condition to be able to obtain unbiased estimates is that the conditional distribution of the imputed variable is the same in the source data and the outcome data:

$$f_{X_{Mis}|Z_{Mis}}(X|Z=z) = f_{X_{Source}|Z_{Source}}(X|Z=z) \tag{1}$$

Weaker conditions than (1) are often sufficient, e.g. that the parameters of $f_{X_{Mis}|Z_{Mis}}$ are identified from the source data or even only a subset of them such as the parameters of $\mathbb{E}(X_{Mis}|Z_{Mis})$ if the outcome model is linear. This makes it desirable to have imputation models that can be adapted to the outcome model, so that they only rely on the necessary assumptions, and that are able to impose restrictions or use additional information if identification of the necessary parameters is not straightforward.

The first factor that determines the imputation bias is selection, since (1) is a "no selection on unobservables" condition that requires the two data sources to be the same in terms of unobservables given the observed variables. The distribution of the observables may differ, allowing stratification or selection on observables. Conditions under which selection on unobservables is likely, potential biases and what can be done in such cases are discussed extensively in the program evaluation literature. Consequently, I focus on an issue with (1) that is more prevalent in the imputation case, which is that (1) is unlikely to hold unless all variables measure the same concepts in the source and the outcome data. This is less likely to be a problem if the source data is from the same survey as the outcome data, but may be violated even in such cases if the data is not (conditionally) missing at random. Bollinger and Hirsch (2013) and Meyer, Goerge and Mittag (2013) provide evidence against the assumption that missing values are (conditionally) random. Tarozzi and Deaton (2009) discuss the measurement problem in the context of imputing income measures from a survey to a census, but the problems they line out are similar in other applications. The assumption that the variables measure the same concept could be violated for various reasons such as whether a given value means the same in the two populations (e.g. subjective assessments of health or well-being) or because there were differences in the way it was collected (e.g. people may respond differently to a question when asked in person or by mail). This requirement makes conditioning variables that are easy to measure prefer-

able. For the variable that is imputed, it should be kept in mind that the imputations always measure what they measured in the source data. Thus, when predicting income based on assets as in Tarozzi and Deaton (2009), the imputations are a measure of how much income a *survey respondent* with these assets would be expected to have. They discuss this problem, conditions under which it occurs and its consequences in detail. This stresses the importance of good source data, since it determines whether the necessary conditions for imputation hold and what the imputations actually measure.

In summary, imputing a variable that is entirely missing using data from a different source is always be worth thinking about if the missing variable is of interest or omitting it causes bias. If (1) or a weaker sufficient condition for unbiased imputation holds, imputation will certainly be beneficial. Even if it does not hold, one may still impute the variable if one believes that the omitted variable bias is worse than the bias from imputation. The selection process into the source data, whether the conditioning variables Z measure the same in the two datasets and whether the concept measured by the imputed variable in the source data is appropriate for the outcome model are the key determinants of this decision. This emphasizes two desirable features of imputation methods: First, the ability to incorporate a large conditioning set in order to make (1) hold and second, to be able to impose additional constraints, both on the estimation of the imputation model and when imputing, if the variables or the distributions differ in a known way (e.g. if the variance of the missing values is known).

2.2 Imputing Missing Values

If some values of a variable are missing, e.g. due to survey non-response, imputations can be done based on the complete cases, which mitigates many of the problems above. Common data sets often already include imputed values. Contrary to the case where the variable is missing entirely, the researcher also has the option to drop the observations with missing values and use only X_{Obs}. Whether it is better to impute or only use X_{Obs} depends on two key factors: whether the missing data mechanism is ignorable and whether the imputations contain information from outside the outcome model. A missing data mechanism is ignorable if it does not bias the estimates. Several sufficient conditions have been derived (e.g. Heitjan and Rubin, 1991; Little and Rubin, 2002), because ignorability is often hard to assess in practice. The discussion here is theoretical and thus kept in terms of ignorability, in practice one may focus on one of the sufficient conditions instead. I consider imputations to contain information from outside the model if they include information that is relevant to the outcome model, but not predicted by the other covariates in the outcome model. That is, the covariates in the outcome model do not perfectly predict the imputed values[1]. The imputations

[1]For simplicity, this definition abstracts from cases in which some of the information in the covariates is not used by the model. For example, linear regressions do not use information from higher order moments. If the other covariates perfectly

could contain outside information for example because they are based on variables that are not used in the outcome model, impose constraints that are not imposed in the outcome model or make use of additional data. I argue below that imputations should mainly be used if ignorability fails, but the imputations contain information from outside the model. Under ignorability one at most gains efficiency from using imputations at the risk of introducing biases. If the imputations do not include information from outside the model, there is little reason to believe that they would reduce biases.

First, consider the case when the missing data mechanism is ignorable. In this case, the complete cases yield unbiased estimates of the model parameters by definition, so only efficiency can be gained from imputing the missing values. This is analogous to the case above in which there is no omitted variable bias, but for partially missing variables imputations are often used even though ignorability is assumed, so it is important to know the conditions under which this may make sense. If the imputations do not contain information from outside the model, imputations of a dependent variable should never be used. Actual values of the imputed variable are perfectly predicted by the other variables in the model, so they just reproduce the exact relationship among (Y_{obs}, X_{Obs}). Consequently, no efficiency gains are possible. Imputations for an independent variable may be useful if efficiency is a crucial issue, the imputation method avoids casewise deletion and one believes that the additional cases contain a lot of information. There cannot be any gains in efficiency from the actual values of an imputed independent variable, since they are by definition perfectly predicted by the other conditioning variables. However, they allow the researcher to include the observations that had missing values. Thereby, information on the other variables in X and Y is used, which may increase efficiency. Consequently, gains in efficiency are possible in some cases, but in many cases it is preferable to forgo these gains to avoid potential bias from the imputations by analyzing only the complete cases. Imputations of a dependent variable should never be used in this situation.

If the imputations include information from outside the model, efficiency is still the only potential gain under ignorability, because there is no bias. However, the imputed values now also provide information about the variable that is missing, which may justify using an imputed dependent variable and adds a second source of potential efficiency gains when imputing an independent variable. The trade-off with potential biases remains, so unless efficiency is a crucial issue, dropping the observations with missing values should be preferable. If efficiency is important, one could consider using the imputed values if the conditions above hold or if the imputations contain a lot of new information. In both cases, the researcher has two unbiased point estimates that should be the same, so a Hausman (1978) test can be used to test for the presence of biases.

predict the conditional mean of the imputed variable, but not its higher order moments, the imputations do not contain information from outside the model.

On the other hand, if the missing data mechanism is not ignorable, using only complete cases analysis yields biased estimates. If the imputations do not contain any information from outside the model, the same problems as in the first case arise, because the imputed values are perfectly predicted by the other variables and thus do not add any information to the biased model. In consequence, it does not make sense to impute a dependent variable, but imputations of an independent variable that avoid casewise deletion may be beneficial. Non-ignorability does not necessarily imply that X_{Imp} is a biased predictor of X_{Mis} (but some of the sufficient conditions imply it), but the fact that the missing data is not ignorable strongly suggests that the imputations cannot solve the selection problem entirely and are likely to cause bias themselves. Consequently, the researcher has to hope that two wrongs make a right (or less wrong), which should not be done without careful analysis of the likely directions and sizes of the biases. In general, the benefit of the imputations is to improve the selection problem by including the observations that were selected out due to missing data. The downside is that the imputation error, η, is related to the error term of the outcome model (otherwise the missing data would be ignorable) and is unlikely to be independent of X_{Imp}, which introduces a form of non-classical measurement error. Therefore, the researcher faces a trade-off between bias due to selection and bias from non-classical measurement error. To illustrate this, consider the case of a linear model in which the researcher uses $\tilde{X}$ instead of X and thus estimates β by

$$\hat{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y$$

Using the true model ($Y = X\beta + \varepsilon$), $X = \tilde{X} - \eta$ and taking expectations gives

$$\mathbb{E}(\hat{\beta}) = \beta + \mathbb{E}\left[(\tilde{X}'\tilde{X})^{-1}\tilde{X}'(-\eta\beta + \varepsilon)\right]$$

There is no bias if $\mathbb{E}(\tilde{X}'(-\eta\beta + \varepsilon)) = 0$. This term can be decomposed as

$$\tilde{X}'(-\eta\beta + \varepsilon) = \underbrace{[X'_{Imp}X_{Mis} - X'_{Imp}X_{Imp}]\beta}_{\text{New Bias I}} + \underbrace{X'_{Obs}\varepsilon_{Obs}}_{\text{Selection Bias}} + \underbrace{X'_{Mis}\varepsilon_{Mis}}_{\text{Cancels Selection Bias}} + \underbrace{\eta'\varepsilon}_{\text{New Bias II}} \tag{2}$$

where the second term on the right is the source of the bias in the complete case analysis. The benefit of the imputations is that this term cancels with the next term, since $\mathbb{E}[X'_{Obs}\varepsilon_{Obs} + X'_{Mis}\varepsilon_{Mis}] = \mathbb{E}[X'\varepsilon]$, which is 0 by assumption. If this outweighs the two new biases from imputation, the imputations improve the estimates and should be used. The bias from imputations is given by the expectation of (2) premultiplied

by $(\tilde{X}'\tilde{X})^{-1}$:

$$\mathbb{E}(\hat{\beta}-\beta)=\mathbb{E}\left[(\tilde{X}'\tilde{X})^{-1}\left[(X'_{Imp}X_{Mis}-X'_{Imp}X_{Imp})\beta+(X_{Imp}-X_{Mis})'\varepsilon_{Mis}\right]\right] \tag{3}$$

The imputations should be used if this bias is less than the selection bias that arises from using only complete cases, $\mathbb{E}[(X'_{Obs}X_{Obs})^{-1}X'_{Obs}\varepsilon_{Obs}]$. The second term in the inner brackets of (3) arises from a relation between the prediction error of the imputed values and the error term. The better the imputations are at reproducing the relation between X_{Mis} and ε, the lower one should expect this bias to be. Because non-response is not ignorable (so the relation between X_{Mis} and ε is not the same as between X_{Obs} and ε) and the imputations do not use additional information, it is unclear why the imputations would be good at this so that improvements are likely to be by chance. Similarly, the size of the first term is determined by how well the imputations reproduce the covariance of X_{Mis}. This term may be zero if they reproduce it perfectly, but as was pointed out above, it is likely that X_{Imp} is a biased predictor of X_{Mis}, so in most cases one should expect this term to be non-zero as well. This trade-off is more complicated in non-linear models since additional biases from higher moments are likely (see e.g. Carroll et al., 2006), but how well the imputations reproduce the two covariances is still the main determinant of the trade off. Unless information on the likely size of the three bias terms is available, using imputations under these circumstances thus is a gamble that could increase or reduce the bias. So unless there is reason to believe that the imputations are good at reproducing $X'_{Mis}X_{Mis}$ and $X'_{Mis}\varepsilon$, they should not be used.

As in the ignorable case, the arguments for imputations are better if they contain information from outside the model. Such additional information may enable the researcher to create imputations that reproduce the covariance of X_{Mis} and make the imputation error orthogonal to ε. Thus, imputations based on additional information can yield unbiased estimates while complete case analysis does not, so imputation has the largest potential value when missing data is not ignorable and information from outside the model is available for the imputations. However, additional information may also introduce additional biases. So whether one wants to use imputations still depends on what kind of information was used and how, as it has to make the terms in (3) small.

In summary, whether gains from imputations are possible if a variable is partly missing depends on the assumptions regarding the missing data mechanism and the informational content of the imputations. If the mechanism is ignorable, efficiency has to be an important issue and strong arguments regarding the benefits of the incomplete cases have to be made to justify imputation. If the mechanism is not ignorable, the decision depends on the informational content of the imputations. If they are based on the same information as the outcome model, improvements are possible, but require a lot of faith in the predictive power of the

imputations that is hard to justify based on the way they are done. The main case for imputing a partly missing variable arises when the missing data mechanism is not ignorable and outside information is used, but one should carefully assess which information is used to make sure it actually solves the problems described above. If an entire variable is missing, imputation is beneficial if the bias it induces is less than the omitted variable bias. Whether this is the case mainly depends on the quality of the source data, in particular whether the imputed variable and the conditioning variables measure the same concepts in the source and outcome data and whether the relation between them is the same in the two data sets. Thus, the greater the similarities between the source and the outcome data and the simpler the concepts being measured, the more promising will imputation be.

The conditions under which imputations can be beneficial have several implications for the imputation methods. First, they need to be able to incorporate a lot of information in order to justify (1), improve efficiency or make (3) small. Second, they need to be good at reproducing the association between the imputed variable and the observable and unobservable components of the outcome model. Third, they should be able to incorporate knowledge about the differences between the source and the outcome data, such as information on selection models, etc. Finally, they need to be transparent so that the researcher knows which information was used to impute values and can judge the likely consequences, because this determines whether imputations should be used at all.

3 Problems with Imputations

This section discusses common problems that are caused by using imputation methods. The first part lines out how to choose the conditioning variables and the biases that result from a bad choice, the second part discusses problems that arise when working with imputed values and the third part shows that the relation between the outcome and the imputation model is important. Finally, I analyze how this should influence the decision to use imputed values or not and which features of imputation models can help avoid these problems. Two important conclusions are that more information needs to be made available about imputation in common data sets in order to address these problems and that the current division of labor in which providers of data sets impute values and researchers use these values is problematic because it amplifies the problems discussed here.

3.1 The Imputation Model: Choosing the Conditioning Set

The choice of conditioning set refers to the choice of variables in Z and is important for two reasons. First, including variables in Z that are not included in X is a way to make the imputations depend on

information from outside the outcome model. The previous section has shown how such information can improve estimates. Second, a badly chosen conditioning set can make the imputations endogenous and thereby lead to bias, which is discussed in this section. I first discuss the consequences of omitting relevant variables from the conditioning set and then show that the conditioning set can also contain too many variables. Hirsch and Schumacher (2004) and Bollinger and Hirsch (2006) have derived formulas for the bias in linear model when the dependent variable is imputed and some independent variables are omitted from the conditioning set for the imputations. They use the CPS Outgoing Rotation Group (ORG) to show that these biases can be substantial. The key problem is that conditional on the variables that are included in the conditioning set, the imputed variable is independent of any other variable. If the imputed variable is a dependent variable, the coefficients on the omitted variables are 0 if only imputed observations are used. As long as Y_{Obs} is not empty, the imputed observations are mixed with complete observations. For most common estimators[2], this implies that the estimates are a weighted average of $\hat{Q}(Y_{imp}, X_{Imp})$ and $\hat{Q}(Y_{Obs}, X_{Obs})$. If $\hat{Q}(Y_{Obs}, X_{Obs})$ is consistent, because missing data is ignorable as Bollinger and Hirsch (2006) assume, the coefficient on the imputed variable will be attenuated. This may not hold if $\hat{Q}(Y_{Obs}, X_{Obs})$ is upwards biased or has the wrong sign. Bollinger and Hirsch (2006) also consider the problem of imperfect matching that arises if a variable is included in the conditioning set, but its functional form is not correct. This is frequently the case in practice, since imputations are often done within cells formed by categorical variables. If the outcome model contains continuous variables or finer levels of the categorical variables in Z, the conditioning set is misspecified which results in a similar bias.

Table 1 illustrates these problems by regressing log-earnings on observable characteristics separately for the imputed and non-imputed observations of the 2010 CPS ASEC. Columns 2 and 3 report the coefficients or, for rows with multiple coefficients, the average absolute deviation of the coefficients from 0. Bollinger and Hirsch (2006) present a similar table for the CPS-ORG, but their results do not apply directly to the CPS ASEC, because the imputation procedure in the CPS ASEC is a sequential hot deck (see e.g. David et al., 1986; Lillard, Smith and Welch, 1986, for descriptions). It uses different conditioning sets for different patterns of missing values and reduces the conditioning set if no match is found. Thus, sorting variables into match variables and imperfect match variables is not cleanly possible, because a variable may be a match variable for a part of the sample, an imperfect match variable for another part and a non-match variable for the remainder of the sample. In the first pass, gender, work experience, class of worker, marital status, occupation, age, family relationship and farm status are match variables and education, Census division, metro size and race are imperfect match variables. However, variables are sequentially dropped from the conditioning set for the observations for which the donor cell is empty in the first pass. Nonetheless, the

[2] A sufficient condition is that the objective function is unimodal.

Table 1: Consequences of an Incomplete Conditioning Set

	Number of Coef.	Coefficient/Average Absolute Deviation		p-Value Joint Equality
		Reported	Imputed	
Dummy Variables				
All Complete Match Dummies	36	0.3769	0.4683	0.00
All Partial Match Dummies	32	0.2839	0.1764	0.00
All Non-Match Dummies	24	0.1831	0.0778	0.00
Female	1	0.1805	0.2533	0.00
Work Experience	11	0.9150	1.1200	0.00
Class of Worker	5	0.0893	0.0680	0.89
Married	1	0.0809	0.0588	0.38
Major Occupation	9	0.2576	0.3040	0.00
Age Categories	4	0.0495	0.1336	0.01
Family Relationship	4	0.0483	0.1096	0.00
Farm Status	1	-0.0852	-0.1792	0.08
Educational Attainment	15	0.4764	0.3281	0.06
Census Division	8	0.1174	0.0217	0.00
Metro Size	6	0.1345	0.0680	0.00
Race Categories	3	0.0643	0.0474	0.07
Hispanic	1	0.0869	0.0399	0.08
Major Industry	12	0.2181	0.0939	0.00
Disabled Person in HH	1	-0.1371	-0.0451	0.00
Union Member	1	0.1395	0.0665	0.06
Foreign Born	1	-0.1009	-0.0691	0.19
Works Part time	1	-0.1390	-0.0301	0.00
Works for Non-Profit	1	-0.0748	-0.0226	0.10
Continuous Variables				
Weekly hours worked	1	0.0203	0.0095	0.00
# of persons under 18 in HH	1	0.0036	-0.0260	0.00
Potential Experience	1	0.0519	0.0170	0.00
...incl. higher order terms	4	0.0135	0.0044	0.00

Notes: N: 70,643 (column 2), 15,836 (column 3). Sample restricted to individuals 18 or older with positive earnings that are no full-time students, in the army or live in group quarters. The first 8 sets of dummies (Female-Farm Status) are considered complete match variables, the following 4 sets of dummies (Educational Attainment-Race Categories) partial match variables, the remainder are non-match variables.

results conform to expectation: non-match variables and partial match variables are attenuated. Most notably, geographic information, industry and how much the individual worked is only weakly related to earnings. Including the imputed observations seriously biases such analyses, particularly since almost 20% of the sample is imputed. If non-response is ignorable, as the imputation procedure assumes, the coefficients in the two columns should be equal. Column 4 contains the p-value of a test that the coefficients are jointly equal, which shows that if it is indeed ignorable, there is a serious problem with the imputations. If non-response is not ignorable, the coefficients need not be equal and it could be argued that the differences between the two columns are due to differences between respondents and non-respondents and therefore desirable.

Since the imputations are only based on the complete cases and the differences conform to what one would expect when using badly conditioned imputations, the more likely explanation is that the coefficients in column 4 are biased because of the imputation procedure. Obviously, no one would use only the imputed values to run a regression, but table 1 shows that including imputations based on an imperfect conditioning set introduces a substantial amount of noise that will bias the coefficients.

Bollinger and Hirsch (2006) focus on the coefficients on the variables that are omitted from the conditioning set, but table 1 underlines that the coefficients on the variables that are included in the conditioning set are also affected. Conditional on the other covariates, the imputed values contain only noise. Thus, if the entire variable were imputed, the bias for the variables that are included in the conditioning set would be equal (in expectation) to the bias from omitting the imputed variable from X. Imputing only a subset of the observations mitigates the problem and including variables in the conditioning set that are not included in the model can also help.

Bias also arises when an incomplete conditioning set is used for an independent variable. Bollinger and Hirsch (2006) do not discuss this case, but their analogy to measurement error extends to it immediately. Let (Z, W) be the appropriate conditioning set, so that W is the set of variables that are mistakenly excluded from the conditioning set. Then the imputed values can be written as

$$\tilde{X} = X + \eta = X + f(W) + \nu \tag{4}$$

where $f(W)$ is the part of X_{Mis} that is predicted by W, but not by the variables that are included in the conditioning set. ν is an error term that is unrelated to X by the assumption that (Z, W) is a sufficient conditioning set. $f(W)$ is related to X by construction, so (4) shows that omitting relevant variables from the conditioning set when imputing an independent variable can be analyzed like non-classical measurement error. The nature of the error depends on the imputation model and the bias depends on the outcome model, so no general results are available. Even if the conditioning set is well chosen, ν causes classical measurement error which is discussed further below. The results so far are mainly for linear models, but since the problem can be seen as a problem of measurement error, the consequences are likely to be worse, but the determinants similar in non-linear models (see e.g. Carroll et al., 2006).

These results seem to underline the common guideline that the conditioning set for imputations should contain as many variables as possible. Steuerle-Schofield et al. (2012) show (for the related case of plausible values) that this is not true, because including Y in the conditioning set induces correlation of the imputed variable with the error term and thereby causes bias. More generally, the instrumental variable assumptions have to hold for the variables in the conditioning set: they have to be related to the missing variable, but

need to be unrelated to the error term of the true outcome model. If the second assumption does not hold, X_{Imp} is related to the error term since it is a function of Z and thereby causes bias due to endogeneity. Thus, variables that are excluded from the outcome model should only be used in the imputation model if the exclusion restriction is likely to hold. Consequently, the conditioning set should include all independent variables of the outcome model, include any additional variables if and only if there are good reasons to believe the exclusion restriction is valid and never include any of the dependent variables.

In conclusion, the conditioning set is crucial for the performance of the imputations and thereby a key factor in the decision whether imputations should be used. Therefore, it is important for the researcher to know the conditioning set of the imputations. This is basically impossible for the imputations in most data sets. The information provided about the imputation procedure is usually sparse and often does not include the variables that were actually used and how they are defined. Additionally, the ideal conditioning set depends on the variables that are included and excluded from the model, so different applications require different conditioning sets. Researchers should therefore be able to chose a conditioning set that is appropriate for their model, which means they either have to do the imputations themselves or should be able to chose from multiple imputations. In terms of imputation methods, this reiterates the point that being able to condition flexibly on many variables is an advantage. It also underlines the importance of having specification tests for the imputation model to avoid misspecification of the conditioning set. Finally, it is an advantage if the method makes it easy to adjust the conditioning set or to let the researcher choose from different conditioning sets.

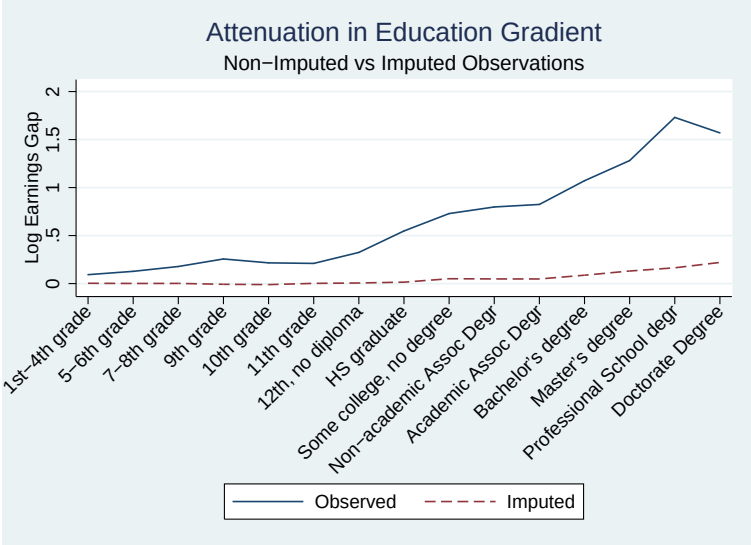
3.2 Working with Imputed Values

If imputed values are used, they are often treated like complete observations in the outcome model. This is wrong for two reasons: First, the imputed values are predicted and therefore only a proxy for the true values, which causes bias from measurement error. Second, they introduce uncertainty beyond the sampling variation to the model, which should be accounted for when estimating the variance. This section illustrates the consequences of treating imputed observations like complete observations and argues that more information about the imputation process needs to be available to address this issue.

I assume that the conditioning set is well-specified, so the imputed variable only suffers from classical measurement error (CME) as equation (4) shows. This does not cause bias if a dependent variable is imputed in linear models. The consequences of CME on an independent variable are known for many models, most notably the result that in linear models the coefficient on the mismeasured variable is attenuated, but other coefficients can be biased upwards or downwards. Figure 1 illustrates this problem by plotting the returns

to education when using imputed and non-imputed observations. To separate the problem of measurement error from a badly chosen conditioning set, the graph is based on an MC study in which the 2010 CPS ASEC sample is randomly split in halves and education is imputed for one half by a hot deck procedure using the observed values of the other half. The conditioning set is chosen in line with the outcome model and non-response is ignorable by construction, so the difference between the coefficients is entirely due to prediction error. It makes the imputations sharply understate the earnings gap between higher educated people and those with less than 1 year of schooling, which is the omitted category. The size of the bias depends on

Figure 1: Attenuation Bias from Imputation



Note: See table 1 for information on the sample, which was randomly split in halves 1000 times and one half was imputed using a hot deck imputation with cells formed by dummies for married, part time, gender, census region and whether the respondent is white. Reported are the average coefficients on education dummies with "less than 1st grade" omitted, the dependent variable is log earnings and the regression also controls for the variables used in the hot deck and a quartic in potential experience.

the signal-to-noise ratio and is larger the lower this ratio is. Thus, the lower the variance of the prediction error η, the lower the bias, so the more explanatory power the imputation model has, the better. Figure 1 does not attempt to replicate the CPS hot deck, but Black, Sanders and Taylor (2003) provide evidence that imputed values for education contain a substantial amount of noise with error rates consistently above 80%. This creates potentially large biases that are avoidable by treating the imputed values appropriately. One way to do so is to estimate the signal-to-noise ratio from the imputation model and use it to correct the OLS estimates. Another option is to employ a consistent estimator, such as the ones below that integrate over the imputed values[3].

Even if the imputations do not cause any bias, one needs to account for the additional uncertainty that is introduced by the prediction error and any parameters that were estimated in the imputation model. Ignoring

[3]Note that multiple imputation, despite integrating over the imputed values, usually only corrects the standard errors because it assumes that there is no bias (see e.g. Rubin, 1996, equation 2.6). However, using multiple imputations as in section 5 can solve this problem.

this uncertainty leads to biased variance estimates. Shao and Sitter (1996) analyze the implications in theory, Andridge and Little (2010) focus on hot deck methods and illustrate the consequences. Both papers conclude that no analytic variance estimators exist for most common imputation methods, particularly for hot deck methods, but some simulation methods have been proposed. By far the most common approach is to use multiple imputations (Rubin, 1987), which works if the imputation method is proper (see Rubin, 1987, p. 118-119). Although most imputation methods are not proper, Andridge and Little (2010) provide evidence that it improves coverage rates compared to "naive" variance estimates. Consistent variance estimates can be obtained through jackknife (Rao and Shao, 1992; Rao and Sitter, 1995) or, more generally, bootstrap methods (Shao and Sitter, 1996), but they are rarely used in practice. One reason for this is that all simulation methods, including multiple imputations, require repeatedly imputing the dataset, which is computationally burdensome. More importantly, it is often impossible to reproduce the original imputation procedure. It requires the researcher to know which observations are imputed (which is now provided in most data sets), how they were imputed (which is usually not provided as I argued above) and to have access to the source data (which is often impossible because imputations are drawn from previous surveys or the restricted-use data).

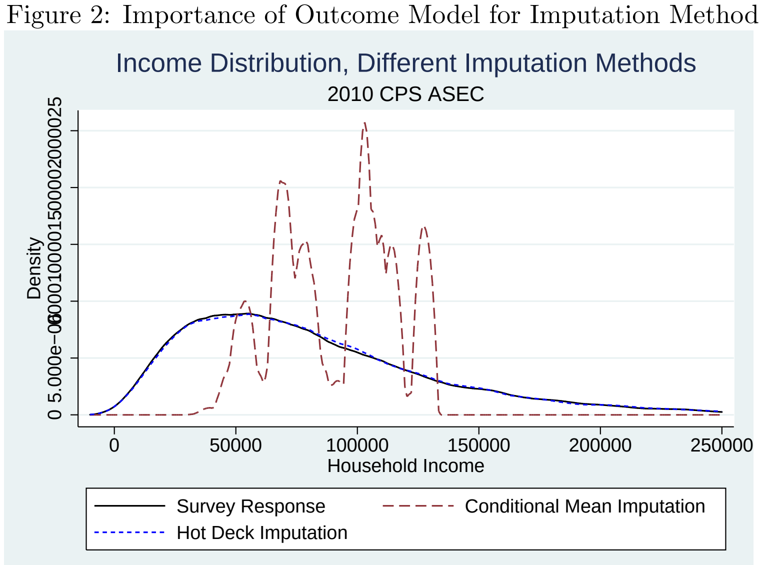
In conclusion, treating imputed observations like complete observations biases coefficients and standard errors even with a correctly specified conditioning set. The size of the bias depends on the outcome model and the imputation method used, but usually decreases with the precision of the imputations. Correcting the bias or making a sound judgment whether the bias from using the imputations or using only complete cases is larger requires more information on the imputation procedure than is commonly provided, particularly on the degree of prediction error. Additional information is also required to obtain consistent standard errors. Unfortunately, the most common imputation method, the hot deck, suffers from severe problems on both accounts, with little chances of fixing them. Other methods, such as some of those discussed below, either do not face these problems at all, can be expected to perform better or the problems are easier to fix. The problems discussed in this section make methods attractive that avoid filling in values, e.g. by integration, because they are often unbiased. To avoid inconsistent variance estimates, imputation methods that provide analytic formulas for the variance, are proper or are easy to replicate in simulation methods should be used.

3.3 Congruence of Imputation Method and Outcome Model

It is well known that the performance of a particular imputation procedure depends on the outcome model used. For example, classical measurement error is often not an issue when imputing a dependent variable in a linear model and is much easier to correct in linear models than in non-linear ones. Most imputation

procedures are better at reproducing certain features of the data than others, so one needs to choose an imputation method that reproduces the characteristics that are important to the outcome model well. On the other hand, outcome models are often robust to some data problems which allows imputation methods to perform well that would induce bias in other cases. For example, conditional mean imputation is good at reproducing the relation between the average value of the imputed variable and its predictors, but bad at reproducing its conditional variance. Hot deck methods, on the other hand, tend to reproduce marginal distributions well, but often mute the association between multiple variables. Consequently, hot deck methods fare poorly for independent variables in linear regressions as figure 1 shows, while conditional mean imputation is known to be unbiased and efficient (Schafer and Schenker, 2000), because linear models are robust to misspecification that does not affect the conditional mean. Figure 2 demonstrates that the reverse is true when the outcome model is a marginal distribution.

Figure 2: Importance of Outcome Model for Imputation Method



Note: See table 1 for information on the sample. The sample was split in halves and one half was imputed based on the other using age, education, female and white dummies as the conditioning set for both imputations.

Overall, this means that different imputation methods should be used depending on factors such as whether the imputed variable is a dependent or an independent variable and whether the outcome model is univariate or multivariate. Currently, most data sets include imputed observations and the researcher needs to check whether they are done in a way that works well in the outcome model. The requirement for congruence of imputation method and outcome model makes such out of the box solutions unattractive and provides another argument to either do the imputations oneself or to provide multiple sets of imputations for different types of outcome models. Another option would be to use imputation methods that work well in many circumstances and are not only ideal for a specific purpose.

3.4 Implications

This part discussed some problems with imputations that cause bias in common applications. These results are informative about the conditions under which imputations should be used as well as how they should be done and the properties that make a good imputation method. If the question is only whether or not to use a particular set of imputations, the potential biases in this section need to be weighted against the benefits in section 2. Three key factors that are likely to make the biases small have been lined out. First, the conditioning set of the imputations should contain all independent variables of the outcome model and besides them only variables that could be used as instrumental variables in the outcome models. As long as the exclusion restriction is plausible, having more such variables in the conditioning set favors using the imputations. Second, the estimator of the outcome model should be able to take the prediction error of the imputed values into account. If it does not, the balance shifts against the imputations with the sensitivity of the outcome model to measurement error and the variance of the prediction error. Third, there should be a good match between the estimator of the outcome model and the imputation method. The better the imputation method reproduces the features of the data that are actually used by the estimator of the outcome model and the more robust the outcome model is to shortcomings of the imputation procedure, the better the case for using imputations.

A reoccurring theme has been that a given set of imputations may work very well under some circumstances and very poorly under others. This raises the question whether the current practice of providing data sets that include imputations is ideal. Having only one draw from the imputation model makes it very hard to avoid bias and obtain correct standard error even if the imputations method and the outcome model are a good match. Even worse, it is often impossible to determine whether there is a good match and how to avoid the problems because not enough information on the imputation method is provided. Ideally, the providers of the data set would provide enough information to reproduce the imputation procedure, which would not only allow researchers to assess its properties, but also enable them to obtain correct standard error and avoid biases in many cases. Unfortunately, this is very complicated or even impossible for most commonly used hot deck imputations.

Instead of a set of imputed observations, a model from which the imputations can be generated could be provided or the researcher could develop the imputations specifically for the outcome model at hand. Developing an imputation model for the application at hand allows the researcher to choose an imputation model that is ideal for this specific purpose. Besides the inconvenience, the main downside of this is that the provider of the data set may have additional information (or expertise regarding imputations) that could be used to make the imputations better. For example, the American Community Survey (ACS) uses a

geosort (U.S. Census Bureau, 2009) in the imputation procedure that cannot be replicated with the public use data because it is based on very fine geographic information. In such cases, a good solution may be to provide a model from which the researcher can draw imputations, which enables the researcher to replicate the procedure and take the "private information" into account. While this does not work for all imputation methods and may require parametric constraints, it has the additional advantage that the researcher knows where the imputations are from and can thus judge whether they are likely to improve things. Depending on the model, the researcher may even be able to impose additional constraints to make the imputations fit the application better. In most cases, all that is required to draw imputations are a small number of parameters, so the data provider could even provide multiple models for different outcome models or using different conditioning sets. Little (2012) discusses further problems, particularly more conceptual ones, that arise from the current practice of providing imputed data sets.

Finally, the problems in this section underline some properties of imputation methods that are desirable. It reinforces the idea that the imputation model should be able to condition on a lot of information in a flexible way. This is important both to be able to use the correct conditioning set and to increase predictive power in order to mitigate the bias due to measurement error. In order to come up with a good conditioning set, it is important to assess the fit and validity of the imputation model, which is easier if there are specification tests for the imputation model. Imputation methods that avoid predicting single values by using analytic solutions or integration often avoid the bias from measurement error altogether. Ideally, the imputation method would allow the researcher to calculate the standard errors analytically, but if this is not possible it is desirable if SEs are reasonably easy to obtain by replication methods, which requires the imputation method to be proper as used by Rubin (1987) or easy to replicate. Finally, the ideal specification of the imputation method depends on the type of outcome model as well as the variables used. If the provider of the data creates the imputations, the former makes imputation methods attractive that work reasonably well under many different circumstances. A solution to both problems is to provide more than one imputation from different models, which is only feasible if the models can be summarized in terms of a few parameters. If this is not possible, the dependence on the outcome model induces a preference for methods that allow imposing constraints on the imputations or other ways of adjusting the conditioning set "ex post".

4 Common Imputation Methods

This section briefly reviews in how far common missing data methods fulfill the criteria that were found to be important above. Are they good at reproducing the marginal distribution or relations between multiple variables from the source data? Can they incorporate additional constraints, such as a selection model or

known features of X_{Mis} if the source data is not ideal? How do they handle prediction error and how can correct standard error be obtained? Can they be extended to impute multiple variables and with which outcome models are they likely to work well?

4.1 Hot Deck Methods

Hot deck methods form donor pools for observations with missing data and a randomly drawn value from that pool is assigned to the observation with missing data. Most large economic surveys (e.g. ACS, CPS, SIPP) use "classic" hot deck methods, in which the donor pools are defined by the cells formed by categorical variables. Several variants are discussed in Andridge and Little (2010), who also analyze the performance and problems of common hot deck methods. A method that has been found to perform well in practice is predictive means matching (PMM), which forms donor pools based on the distance between the predicted means from a regression of the missing variable on some covariates. This is a form of matching in which the missing values are replaced by a draw from observations with similar predicted values of the missing variable.

Hot decks only impute values from the original data, which makes them very good at reproducing marginal distributions including idiosyncratic features of the data that would be "smoothed out" by parametric methods. This advantage turns into a downside if the source data suffers from problems and one would like to impose constraints or model selection into the source data. Classic hot deck methods cannot include continuous variables in the conditioning set and are limited in the number of categorical variables because the "curse of dimensionality" quickly makes the number of cells large, which results in imputation from empty or thinly populated cells. Thereby, they often fail to capture multivariate relationships beyond basic ones such as univariate statistics within cells. The main advantage of methods like PMM is that they allow more flexible conditioning sets and thereby capture the relations between variables better, but they make more parametric assumptions which are often chosen arbitrarily, because specification tests are not available. However, the main problem of hot deck methods is that it is very hard to work with the imputed values, because we know very little about their theoretical properties (Andridge and Little, 2010). As figure 1 shows, the imputed values cause bias in models that are sensitive to measurement error, which is hard to correct because the relevant parameters are unknown. Correct SEs can only be obtained by bootstrapping with repeated imputation, which requires the source data to be available and is computationally expensive. Hot deck methods do not extend well to multivariate imputation although some attempts have been made (see e.g. Andridge and Little, 2010).

Overall, this makes the hot deck a rather specialized imputation procedure. If the object of interest is

univariate or a univariate population statistic within one of the imputation cells, the hot deck may perform well since it does not make a functional form assumption on the marginal distributions and only uses the original data. However, in multivariate models it is hard to specify and likely to produce bias and inconsistent SEs.

4.2 Re-Weighting Methods

Researchers commonly adjust weights to account for missing data, usually by inverse probability weighting (IPW, see Wooldridge, 2007), i.e. by multiplying the original weights by the inverse of the estimated probability of response and using only the complete cases. The probability is usually estimated by a standard binary choice model, which allow for large and flexible conditioning sets and are simple to implement and specify, because standard specification tests can be used. If necessary, additional constraints can be imposed and Little and Vartivarian (2005) describe which variables should be included in the conditioning set.

The outcome model only includes complete cases, so there is no prediction error and no infeasible values or combinations occur. SEs can be calculated analytically or bootstrapped which is simple in implementation and computation. It can be extended to multivariate patterns of missing data, since one can estimate the probability of an observation missing for any reason and drop all observations with any missing values. However, this may involve dropping many observations and thereby discarding a lot of information if values are missing from multiple variables. Section 2 has shown that one of the main benefits of imputations stems from enabling the researcher to use the incomplete observations, and IPW forgoes this advantage by only using the complete cases in the outcome model. If the complete and incomplete observations differ on unobservables, the weights may be unable to adjust the distributions. In these cases, IPW tends to make things worse, since it exacerbates selection by overweighting the "selected" observations. Consequently, it is primarily attractive in situations in which missing data is conditionally missing at random, so that the value of imputations is relatively low. IPW can obviously not be used if an entire variable is missing.

In summary, re-weighting methods are limited in scope: they only work for partially missing data and should only be applied if one is confident that the data is missing conditionally at random. However, if the data is indeed missing conditionally at random and the share of complete cases is sufficiently large, they have very attractive properties as they are simple to implement and able to incorporate a lot of information in a flexible way.

4.3 Parametric Methods

Parametric methods assume a fully specified model (with some parameters to be estimated) for the variables with missing values. The model can be stochastic if it specifies that the values are drawn from some distribution and is used to either fill in the missing values or express the outcome model in terms of observed values. Parametric methods vary greatly in the complexity of the model, ranging from simple conditional mean or regression based imputations (e.g. David et al., 1986) to models that allow for selection on unobservables in different ways (e.g. Greenlees, Reece and Zieschang, 1982; Durrant and Skinner, 2006) and their performance is likely to depend on the details of the model.

The parametric assumptions allow them to incorporate a lot of variables and make it easy to impose restrictions or incorporate additional information, but they are likely to be biased if the assumptions fail. However, a large array of tests for parametric assumptions exists and many flexible parametric models are available if the parametric assumptions fail. Bias from prediction error can either be avoided, e.g. by estimating the joint model, or corrected as long as the parametric model identifies the variance of the prediction error. SEs can often be obtained analytically (e.g. using the corrections in Murphy and Topel, 1985; Newey and McFadden, 1994) and almost always by re-sampling methods since it is usually easy to repeat the imputation procedure. Multiple variables can be imputed by combining the parametric models into a joint model or full conditional models that sequentially condition on the previously imputed variables. In light of the fact that the imputation model is often estimated by a different person than the outcome model, an important feature of parametric models is that the imputations can be adjusted to the outcome model even after the imputation model is estimated. The person that estimates the outcome model could, for example, constrain the range of the imputed values or adjust parameter estimates based on additional information (e.g. using the methods in Imbens and Lancaster, 1994). Providing a rather general model for the missing values allows generating imputations that work well in a wide range of settings because different imputations can still be chosen by the researcher. For example, a parametric model of the full conditional distribution of a missing variable allows the person who estimates the outcome model to use draws from this model when estimating a marginal distribution or use it to impute conditional means when running a linear model. The imputation model usually only depends on a few parameters, so estimates for different conditioning sets or models could even be made available.

In summary, parametric model can have great practical advantages and desirable theoretical properties, but they need to be specified correctly. This suggests the development of flexible parametric methods and specification tests for the underlying assumptions.

4.4 Semi-Parametric Methods

Some semi-parametric methods that are directly applicable to the problems discussed in this article have been proposed in the data combination and errors-in-variables literature, for example in Chen, Hong and Tamer (2005) and Ichimura and Martinez-Sanchis (2005, 2009). They can be seen as extensions of the parametric models discussed above that relax some or all of the parametric assumptions. The estimators have very desirable properties in theory, but so far very little is known about their performance and feasibility in practice. In theory, they should be able to condition on other variables very flexibly without constraining the marginal distributions beyond the regularity conditions and thereby reproduce both marginal distributions and multivariate relations well. A common problem with this flexibility, however, is that the "curse of dimensionality" may constrain the conditioning set to few variables. Very flexible estimators also often produce erratic behavior in the tails of the distribution, which can be a problem if the outcome model depends on such characteristics. Sieve estimation of the imputation model as in Chen, Hong and Tamer (2005) may be able to solve both problems. There is no bias due to prediction error and consistent standard errors can easily be calculated. While the methods work for a very general class of models, the model for the missing data that is actually estimated depends on the outcome model. Consequently, the same person needs to estimate the imputation and the outcome model, which also requires the researcher to have access to the source data of the imputations. Nonetheless, these estimators relax the parametric assumptions of the methods discussed in 4.3 and thereby avoid one of the key problems of an otherwise very attractive approach to missing data. The next section considers another way to solve this problem, which may often be preferable in practice. However, in applications where the semi-parametric methods are feasible and found to be stable, they should be used, at least to test whether the parametric models are sensitive to their assumptions.

5 Conditional Density Method

The previous section has shown that parametric methods can have several advantages if they are correctly specified. The method proposed in this section extends the existing parametric methods by making weaker parametric assumptions and providing an imputation model that can be used for many different outcome models. The parametric assumptions it imposes are very flexible and can be tested and relaxed arbitrarily to make sure they are justified.

The key idea is to estimate the conditional distribution of the missing data and use this distribution and the observed data to estimate the parameters of the outcome model. The conditional distribution can easily be made available to researchers and allows them to impute the data in ways that work well for their

outcome model of interest. Consequently, the method is particularly advantageous if the imputation model and the outcome model are estimated by different people. Section 5.1 describes the method, section 5.2 discusses its advantages and section 5.3 compare its performance to other imputation methods both when imputing a variable that is missing from the data partially and entirely.

5.1 The Method

The method I describe in this section is based on an insight from the semi-parametric data combination and measurement error literature, but makes a flexible parametric assumption that results in a number of practical advantages. Its consistency follows from the consistency of the semi-parametric estimators and thus holds for a wide range of models. The conditional density method is based on the same idea as, for example, Hsiao (1989), Li (2002) and Chen, Hong and Tamer (2005). For different applications, they show that the law of total probability can be used to express a model that contains unobserved variables X in terms of observed variables Z and the conditional density of the unobserved variable given the observed variables. For example, the density of X can be written as

$$f_X(X) = \int f_{X|Z}(X|Z = z) \cdot f_Z(z)dz \tag{5}$$

and a (non-linear) regression line, $Y = g(X, \beta)$, implies that

$$\mathbb{E}(Y|Z) = \int g(x, \beta) \cdot f_{X|Z}(x|Z)dx$$

In both cases, the expression on the right is in terms of observables only, so it can be used to estimate the distribution or the parameters of interest if the conditional distribution is known. More generally, it has been shown that a large class of models (including ML, GMM and all common regression models) that depend on an unobserved variable X can be estimated, from observables Z if a model-specific function of the conditional distribution of X given Z, $h(f_{X|Z}(X|Z))$, is known or can be estimated (see e.g. Hsiao, 1989; Sepanski and Carroll, 1993; Lee and Sepanski, 1995; Li, 2002; Chen, Hong and Tamer, 2005; Ichimura and Martinez-Sanchis, 2009). The estimators they propose are thus functions of the observables Z, the function $h()$ and potentially other variables or functions, i.e. the estimators can be written as $\hat{Q}(Z, h(f_{X|Z}(X|Z)), ...)$. The function $h()$ depends on the outcome model, but the key idea here is that it is sufficient to have a consistent estimate of $\hat{f}_{X|Z}(X|Z)$ to estimate the models discussed in these papers. By standard arguments, $h()$ can

be replaced by a consistent estimate, $\hat{h}()$ in the proposed estimators so that

$$\hat{Q}(Z, \hat{h}(f_{X|Z}(X|Z)), ...) \tag{6}$$

also is a consistent estimator if $\hat{h}()$ is consistent. Such a consistent estimate can be obtained from a maximum likelihood estimate[4] of the conditional distribution $\hat{f}_{X|Z}(X|Z)$ if the function $h()$ is continuously differentiable, which all the papers above assume. Consequently, any of the models they discuss can be estimated by

$$\hat{Q}(Z, h(\hat{f}_{X|Z}(X|Z)), ...) \tag{7}$$

if a consistent estimate of $f_{X,Z}(X, Z)$ is available. Chen, Hong and Tamer (2005) consider, but do not implement an estimator that is based on a non-parametric estimate of the conditional density. They argue that the estimator they propose, which directly estimates $h(f_{X|Z}(X|Z))$ instead, is computationally simpler. I use a flexible maximum likelihood estimator for the conditional density, which makes computation easier, but the main reason for estimating the density instead of the function of it is that in many cases the practical advantages of this two-step procedure are likely to outweigh the added computational complexity.

The key requirement to obtain consistent estimates is thus to have a consistent estimate of the conditional distribution of X given Z. Since this paper focuses on imputation based on complete observations or another data set, I discuss estimation from an observed sample of (X, Z). Other approaches can be taken for example based on repeated measures (e.g. Schennach, 2004; Bonhomme and Robin, 2010; Abowd and Stinson, forthcoming) or independence assumptions (e.g. Schennach, 2007; Hu and Schennach, 2008; Hu and Ridder, 2012). Most models for scores such as factor models or item response theory implicitly identify the conditional distribution, so the method could be used to work with such scores. I assume that (1) holds, i.e. that $f_{X|Z}(X|Z)$ is the same in the data from which it is estimated and the data for the outcome model. This assumption could be relaxed by modeling selection into the source data or imposing other constraints. I use a flexible maximum likelihood estimator by assuming that $f_{X|Z}(X|Z)$ is a mixture of parametric density functions, such as a mixture of K normal distributions:

$$f_{X|Z}(X|Z) = \sum_{k=1}^{K} w_k \phi(X; Z\beta_k, \sigma_k) \tag{8}$$

Note that if X is partially observed, the conditional distribution is (8) for X_{Mis}, while the other observations are equal to their observed value with probability 1. If X contains variables without missing values, the

[4] Any other estimator that remains consistent under continuous parameter transformations can be used instead of maximum likelihood.

conditional density is of the latter, degenerate type for all observations. An advantage of the maximum likelihood framework is the large number of tests for the parametric restrictions and the ability to relax them further if they are unwarranted. If normal distributions do not fit the data well, mixtures of other parametric families such as truncated normals, t-distributions or Weibull distributions can be used. If there are other idiosyncrasies in the source data such as mass points, they can be added to the model.

Besides the usual specification tests for maximum likelihood, tests that can be used to assess how well the estimated conditional distribution reproduces characteristics of the source data are of particular interest[5]. Whether the conditional mean is well specified can be assessed by the test in Horowitz and Härdle (1994). A simple test whether the conditional distribution reproduces the marginal distribution of X well is to draw a sample of X from the estimated conditional distribution evaluated at Z_{Source} and test whether it is from the same distribution as X_{Source} using a Kolmogorov-Smirnov test (Smirnov, 1939). To take into account how well it reproduces the relation to the conditioning variables, one can test whether X_{Source} is a sample from the estimated conditional distribution, for example using the conditional Kolmogorov test in Andrews (1997) or the test proposed by Rothe and Wied (2013). If the model does not pass the tests, different parametric families can be used or the parametric assumptions can be relaxed by adding more parameters or components to the mixture. The flexibility of this framework combined with powerful specification tests should make it possible to find a parametric model that fits the data sufficiently well and thereby avoid one of the main disadvantages of parametric imputation models. If no satisfactory model can be found, non-parametric estimation using sieve estimators (Chen, 2007) can be used to relax the distributional assumptions entirely, which yields a two-step estimator as proposed in Chen, Hong and Tamer (2005).

With an estimate of the conditional density, the researcher can choose between using (6) and (7) to estimate the model. Following Chen, Hong and Tamer (2005), (6) is likely to be simpler to estimate, but requires the researcher to calculate $\hat{h}(f_{X|Z}(X|Z))$, whereas (7) directly uses the conditional distribution and does not require any intermediate calculations. Consequently, which of the two estimators is preferable likely depends on how easy it is to calculate $\hat{h}(f_{X|Z}(X|Z))$. For example, in a linear regression, $Y = X\beta + \varepsilon$, the function is $h(f_{X|Z}(X|Z)) = \mathbb{E}[X|Z]$, the conditional mean of X given Z. It is often trivial to calculate, either because it is a simple function of the parameters of the conditional density or by simulation. Thus, for linear regressions an estimator based on (6) seems preferable:

$$\hat{\beta} = \left[\mathbb{E}[X|Z]'\mathbb{E}[X|Z]\right]^{-1} \mathbb{E}[X|Z]'Y \tag{9}$$

For other models, $h()$ will often be an integral without known solution. If random draws form the conditional

[5]Programs to estimate a conditional density by flexible parametric methods that implement the tests below are available from my website.

distribution can be simulated it is likely that an estimator based on (7), such as the moment-based estimator Chen, Hong and Tamer (2005) mention, is simpler:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left[S^{-1} \sum_{s=1}^{S} m(X_s, \beta) \right]' W \left[S^{-1} \sum_{s=1}^{S} m(X_s, \beta) \right] \tag{10}$$

where $m(X, \beta)$ are the moment restrictions of the original model, $\{X_s : s = 1, ..., S\}$ are repeated draws of X from $f_{X|Z}(X|Z)$ evaluated at the observed values of Z and W is a weighting matrix. When estimating a marginal density as in (5), the two estimators coincide since $h()$ is the identity function. Taking random draws $\{X_s : s = 1, ..., S\}$ as before yields a random sample from $f_X(X)$, which can then be used to estimate the marginal density by any standard method such as kernel density estimation.

5.2 Comparison to Other Imputation Methods

This section evaluates the conditional density method proposed above on the same criteria as in section 4. The method proposed can be seen as an extension of the parametric methods discussed in section 4.3 or as a restricted version of some of the semi-parametric estimators in section 4.4. While each of these methods can claim superior theoretical properties in some realm (efficiency for the parametric methods, flexibility for the semi-parametric methods), I argue that the method I proposed is likely to work well in many cases and has a lot of practical advantages.

The conditional density method inherits the advantages of parametric models that they are capable of conditioning on many variables because they avoid the "curse of dimensionality". The flexible framework and the availability of powerful specification tests mitigate the corresponding downside that these assumptions may induce bias if they are wrong, because the parametric assumptions are easy to test and can be relaxed. The model can be extended to situations in which the source data is imperfect in the sense that (1) does not hold by a selection model or by invoking other assumptions such as independence assumptions to identify the conditional distribution. Compared to the semi-parametric estimators, using a parametric model for the conditional distribution has the advantage that it avoids erratic behavior where the source data is thin and that it is straightforward to implement: estimation of the conditional density by maximum likelihood can be done with all common statistical programs. In some cases, estimation based on (6) is simple, but if it is not, a simulated estimator based on (7) can easily be implemented. Sampling from mixtures of parametric distributions is simple[6] and the maximization or minimization problem is the same as the original problem as can be seen from the fact that (10) includes the original moment restrictions $m()$ evaluated at the simulated values. Thus, implementation is straightforward in any statistical program that is capable of

[6] A program that does so for all distributions used in this paper is available from my website.

estimating the outcome model if complete data were available. Like multiple imputation, simulation is used to solve an integral, but the conditional density method is still unbiased if the outcome model is sensitive to measurement error, because the objective function takes the imputation procedure into account. For example, the Probit model can be estimated by simulating d draws from the conditional distribution and maximizing the Probit likelihood over the resulting sample of size $d \cdot N$. Consequently, existing estimation routines can be used to estimate the outcome model with imputations, but the default standard errors will be incorrect. Consistent standard errors can be calculated using the results in Murphy and Topel (1985) and Newey and McFadden (1994), because the estimator is a standard parametric two-step estimator. If a semi-parametric estimator is used instead of maximum likelihood to estimate the conditional density, SEs can be obtained as in Ackerberg, Chen and Hahn (2012). The notation above allows for multiple variables with missing values, but estimating a multivariate conditional density may be infeasible in practice. A simple alternative is to estimate conditional densities that can be used sequentially by conditioning on the imputed values of the variables that were already imputed.

Another advantage of the conditional density method is that it facilitates the division of labor in which different people develop the imputation and the outcome model. Some advantages of this division of labor are discussed in Rubin (1996), Little (2012) points out problems with its current practice. He argues for an integrated model-based approach such as Calibrated Bayes or the frequentist approach taken here. One of the main advantage of the division of labor is that it allows the imputation model to incorporate information that is not available to the analyst who estimates the outcome model, as was discussed above. A key advantage of using a parametric model is that all that is needed to produce imputations or work with the conditional density is the vector of estimated parameters and their covariance matrix. These parameter estimates can easily be made available to the researcher. Thus, the analyst of the outcome model does not need access to the source data, which may be confidential such as administrative records or simply not available. Contrary to common imputation methods such as the hot deck, the conditional density method works well with a wide range of outcome models. The main reason for this is that the analyst of the outcome model chooses the second step estimator instead of being given a set of imputations that works well for some models only. Additionally, the flexibility of the estimator makes it good at reproducing both the marginal density and the relation between variables. Finally, the analyst of the outcome model can still impose constraints on the imputation model, such as scaling up the conditional variance or restricting the range of imputations, if the imputation model is not appropriate in the case at hand. This facilitates the division of labor because the choices the data producer make in the imputation model are less dependent on the outcome models the data will actually be used for. However, it is also easy to make more than one imputation model available to the researcher by estimating different conditional densities. This only requires an additional parameter vector

to be made available and not an additional data set and could allow the researcher to choose from multiple conditioning sets.

5.3 Performance

This section applies the conditional density method to two imputation problems: Imputing food stamp amounts in the ACS using the CPS and imputing hours worked for a subsample of the ACS. The main point is to evaluate the performance of the method for different objectives (data combination and non-response) as well as for different outcome models (univariate and multivariate models, imputing a dependent or an independent variable). I compare the performance of the conditional density method to a classic hot deck, since it is the most commonly used imputation procedure, PMM, because it has been found to perform well and conditional mean imputation[7].

The ACS is the only large economic survey that is representative at the sub-state level, but while it contains information on receipt of government programs, it does not contain information on the amounts received from non-cash programs. Consequently, there has been some interest in imputing these amounts based on the information in the CPS in order to be able to use the ACS to compute estimates at the sub-state level, most notably NAS-style poverty rates (e.g. Levitan and Renwick, 2010). I impute food stamp amounts in the 2010 ACS using the 2010 CPS ASEC and analyze how well the imputation methods reproduce the marginal distribution and the coefficients in a regression of food stamp amounts on some household characteristics. The ACS and the CPS ASEC are sampled from the same population, so assuming that the distribution of food stamp amounts and its relation to other variables is the same in the two surveys seems reasonable. This justifies assumption (1), but it also means that any results at the national level should be the same in the two surveys, so the estimates based on imputations should be the same as the estimates from the CPS ASEC. The first step is to estimate the conditional density of yearly food stamp amounts received by households in the 2010 CPS ASEC. A normal distribution truncated on the left with the mean specified as flexible additive function of the covariates fits the data well:

$$f_{X|Z}(X|Z=z) = \left[1 - \Phi(a, z\beta, \sigma)\right]^{-1} \sigma^{-1} \phi(X; z\beta, \sigma)$$

Adding a second mixture component, including additional higher order terms or relaxing the restriction on the conditional variance does not change the results substantively. Table 2 presents the parameters of the conditional density. I use the same conditioning set for the conditional mean imputation and PMM, and

[7]Programs that implement the estimators for the conditional density and the outcome models discussed in this section are available from my website.

Table 2: Conditional Density of Food Stamp Amounts, CPS

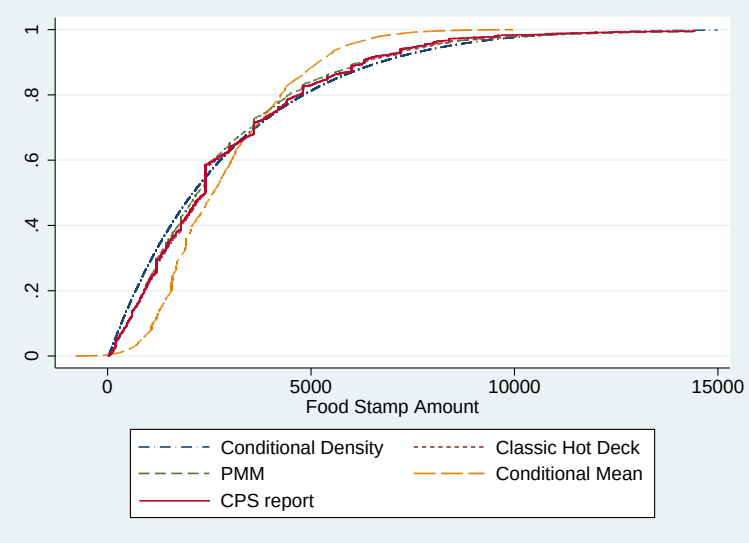
	Coefficient	SE		Coefficient	SE
# of Pers. in HH: 2	3,948***	(670.0)	Earned Income	-0.124***	(0.020)
# of Pers. in HH: 3	6,059***	(886.6)	Earned Income^2	1.6e-06***	(0.000)
# of Pers. in HH: 4	7,662***	(1,049)	Earned Income^3	-0.000**	(0.000)
# of Pers. in HH: 5	8,654***	(1,128)	Any Capital Inc.	-2,114***	(787.3)
# of Pers. in HH: 6	9,298***	(1,265)	Poverty Status	1,643***	(335.8)
# of Pers. in HH $>$ 6	10,273***	(1,331)	Hispanic	-758.7***	(277.6)
# of Children	774.9***	(141.0)	Not a US Citizen	-956.2**	(372.1)
Any Children	853.1*	(436.7)	Constant	-8,112***	(1,729)
Any elderly Pers.	-1,291***	(347.1)			
Any disabled Pers.	-225.7	(233.7)	Sigma	3,682***	(299.8)
Female	1,004***	(247.1)	Left Trunc. Point	16.00***	(0.000)

Notes: N: 6170 households that report food stamp receipt (excluding imputed observations) with income less than \$200k and household head older than 18. Demographic variables are characteristics of the household head. *** p<0.01, ** p<0.05, * p<0.1

cells formed by the number of persons, the number of children and whether an elderly or disabled person is present in the household for the classic hot deck.

The parameters of the conditional density can be used to impute food stamp amounts received in the ACS. Figure 3 compares the CDF[8] in the CPS ASEC to those obtained by different imputation methods in the ACS. Both hot decks and the conditional density method reproduce the CDF quite well, although

Figure 3: Comparing Imputations - CDF of Food Stamp Amounts



Note: Based on the 2010 ACS, see table 2 for the sample definition.

the hot deck methods are better at reproducing the small idiosyncrasies, such as the jumps at some round values. The classic hot deck is slightly closer to the reported distribution in the CPS ASEC. The same

[8]I compare CDFs and not densities because the smoothing required for a density removes the jumps and thereby any difference between CPS reports, the classic hot deck, PMM and the conditional density method.

pattern prevails when looking at moments and quantiles: the three methods perform well, but the hot deck methods perform a bit better. Conditional mean imputation on the other hand fares poorly, because it shrinks the tails of the distribution. Overall, the results conform to expectations based on theory.

In order to assess how well the imputation methods reproduce multivariate relations, I regress the imputed food stamp amounts on some basic household characteristics. I use (6) for the conditional density estimator, i.e. the imputed values are the conditional mean calculated from the parameter estimates of the conditional density. Linear regressions are robust to classical measurement error in the dependent variable, so the problem of bias due to prediction error does not arise for the other imputation methods. For the conditional density method, PMM and conditional mean imputations, the conditioning set contains all covariates in the regression, so these models should be unbiased. The results are presented in table 3 and show that only the

Table 3: Comparing Imputations - Predictors of Food Stamp Amounts, 2010 ACS

	(1)	(2)	(3)	(4)	(5)
	Dependent Variable: Food Stamp Amounts				
	CPS Report	Cond. Density	Classic Hot Deck	PMM	Cond. Mean
# of persons	250.7***	209.5***	331.1***	206.8***	219.5***
	(41.08)	(2.844)	(8.715)	(11.95)	(2.251)
# of children	769.8***	743.7***	670.0***	757.4***	768.6***
	(58.87)	(4.684)	(12.82)	(19.88)	(3.025)
Elderly present	-262.2***	-316.8***	4.076	-381.7***	-293.5***
	(71.35)	(4.789)	(16.10)	(17.57)	(4.656)
Disabled present	213.2***	129.0***	62.99***	13.56	139.8***
	(70.48)	(4.958)	(16.15)	(17.86)	(4.279)
HH head female	438.5***	449.2***	38.77**	436.7***	449.7***
	(71.31)	(4.875)	(16.84)	(18.01)	(4.469)
Constant	905.3***	1,035***	1,057***	1,109***	949.3***
	(96.69)	(6.315)	(22.21)	(23.52)	(6.286)

Notes: N: 6170 (column 1), 127,276 (columns 2-5), see table 2 for the sample definition. Standard Errors in column (2) are adjusted for the two-stage procedure, SEs in columns 3-5 are unadjusted. *** p<0.01, ** p<0.05, * p<0.1

conditional mean imputation does slightly better than the conditional density method. All coefficients from the conditional density method are within two standard deviations from the CPS ASEC estimates and none of them are much further than one standard deviation from the CPS ASEC estimates.

To illustrate how imputations work when a variable is partly missing I impute hours usually worked in the ACS and use it in a regression of earned income on hours worked, education categories, gender, race, marital status, region of residence and a quartic in potential experience. To ensure that observations are missing at random, I randomly split the sample of individuals that worked the entire year in halves and impute weekly hours usually worked in one half based on the other half. I use the same imputation methods as above, in addition to the covariates used in the regression the conditioning set includes the number of persons and own

children in the household. The conditional density method is based on a left truncated normal distribution, changing the specification has some impact on the parameter estimates of the conditional density (available upon request), but does not affect the regression results presented below. The classic hot deck is based on the cells formed by the categorical covariates in the regression. Hours worked is only partly missing and missing values were created randomly, so "truth" is given by a regression that uses the full sample and a complete case analysis is feasible and unbiased. Table 4 presents the results. As expected, both hot deck

Table 4: Imputing an explanatory variable (hours worked)

	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent Variable: Earned Income					
	Full Sample	Complete Cases	Cond. Density	Classic Hot deck	PMM	Cond. Mean
Usual Hours Worked	1,127***	1,118***	1,122***	550.5***	577.3***	1,122***
	(6.798)	(9.732)	(9.714)	(6.082)	(6.194)	(9.714)
Female	-14,392***	-14,394***	-14,410***	-17,013***	-16,910***	-14,410***
	(101.8)	(145.1)	(108.8)	(105.9)	(106.0)	(108.8)
Black	-4,525***	-4,515***	-4,466***	-4,950***	-4,894***	-4,466***
	(145.4)	(208.9)	(148.0)	(149.2)	(149.1)	(148.0)
Married	6,587***	6,622***	6,593***	6,728***	6,796***	6,593***
	(103.5)	(146.4)	(105.2)	(106.1)	(106.0)	(105.2)
High School Degree	9,589***	9,655***	9,585***	10,049***	10,078***	9,585***
	(121.8)	(173.4)	(122.0)	(122.4)	(122.1)	(122.0)
Some College	18,582***	18,738***	18,563***	19,495***	19,564***	18,563***
	(126.0)	(179.6)	(127.1)	(127.3)	(127.0)	(127.2)
College Degree or more	51,224***	51,340***	51,232***	53,226***	53,262***	51,233***
	(168.0)	(238.1)	(171.4)	(172.6)	(172.3)	(171.4)
N	956,321	478,046	956,321	956,321	956,321	956,321

Notes: 2010 ACS sample of individuals 18 and older that are neither in school nor in the army, had positive earnings and worked for 50 weeks or more in the past year. Excludes observations for which hours worked was imputed. The regressions additionally contain 3 region dummies and a quartic in experience, the omitted education category is less than high school. The SEs in (3) are corrected for estimation of the first stage parameters, but the SEs in column 4-6 are not corrected for the imputation procedure. *** p<0.01, ** p<0.05, * p<0.1

methods are biased towards zero due to prediction error. The bias is substantial and partly picked up by the other coefficients which are biased away from zero. PMM does marginally better than the classic hot deck, which is also affected by the exclusion of the continuous variables from the conditioning set, but this bias seems to be small. For the unbiased methods, the main impact of the imputations is to improve efficiency, particularly for the non-imputed variables. This conforms to expectations, since missing values are ignorable by construction, and the conditioning set does not include a lot of information from outside the model. The conditional density method delivers almost identical results as conditional mean imputation. In both cases, the standard errors of the imputed variable are only marginally smaller than in the complete case analysis, which confirms that the conditioning set does not include much information beyond what is already contained in the other covariates. On the other hand, the non-imputed coefficients are estimated almost as precisely

as the coefficients when the variable is completely observed in column 1, which confirms that efficiency gains can be realized for the non-imputed variables.

Overall, these results underline the theoretical considerations from above. In particular, they show that how well an imputation method works depends heavily on the outcome model. For example, the hot deck methods perform well when the object of interest is a marginal density, with the classic hot deck performing better than PMM. In multivariate models, PMM does better than the classic hot deck, which underlines the importance of the conditioning set and the problems of the classic hot deck to incorporate many conditioning variables. Both hot deck methods perform worse than the other imputations in multivariate models, particularly when the imputed variable is an explanatory variable. The conditional density method has the advantage that it performs well in all circumstances tested here, even though the ideal imputation for the case at hand performs better. This suggests that if it is feasible for the researcher to do the imputation model, it may improve estimates. However, as I argued above, it is often infeasible or undesirable for the researcher to do the imputation model, in which case the conditional density method is attractive, because contrary to other methods, it works well in many different applications, is simple to implement and has several practical advantages such as analytic formulas for variance estimation. The last application also underlines that only efficiency can be gained from using imputation methods if missing data is ignorable, but it can lead to considerable biases, which makes its use questionable when not necessary.

6 Conclusion

This paper discusses advantages and problems of missing data methods and evaluates how several methods deal with these issues. Since all of these methods invoke additional assumptions and bear the potential to cause bias, the first question when encountering a missing data problem should always be whether anything can be gained from imputations. If a variable is only partly missing, the main case for imputations arises when the missing data mechanism is not ignorable and the imputations contain information from outside the model that may fix this issue. In other cases, imputations are unlikely to improve estimates unless special arguments about efficiency (under ignorability) or the nature of the imputations can be made. If a variable is missing from the data entirely, imputations can help to reduce omitted variable bias if the source data for the imputations is similar enough to the outcome data. Imputations should only be used if the problems discussed in section 3 are likely to be small relative to their benefits, i.e. if the imputations use a conditioning set that is appropriate for the outcome model, the imputations do not cause bias from prediction error and can be taken into account when estimating the variance and there is a good match between the outcome and the imputation model.

The overview of how current imputation methods address these problems in section 4 reveals many problems, particularly when using hot deck imputations in multivariate models. I argue that making an estimate of the conditional distribution of the missing values available instead of providing predicted values from this distribution would allow researchers to estimate a wide range of models. This would alleviate the bias from the common mismatch between imputation model and outcome model. This conditional density method would also enable researchers to account for the imputed information appropriately and allow data users to adjust imputations provided by the data producer. Thereby, it facilitates the current practice that data providers implement the imputation model, while data users estimate the outcome model. While this division of labor is important because both data providers and users have private information, it is the source of many of the problems with the other imputation models. The conditional density method can improve these problems in practice as the applications in section 5.3 show.

References

Abowd, John M., and Martha Stinson. forthcoming. "Estimating Measurement Error in Annual Job Earnings: A Comparison of Survey and Administrative Data." *The Review of Economics and Statistics*.

Ackerberg, Daniel, Xiaohong Chen, and Jinyong Hahn. 2012. "A Practical Asymptotic Variance Estimator for Two-Step Semiparametric Estimators." *The Review of Economics and Statistics*, 94(2): 481–498.

Andrews, Donald W. K. 1997. "A Conditional Kolmogorov Test." *Econometrica*, 65(5): 1097–1128.

Andridge, Rebecca R., and Roderick J. A. Little. 2010. "A Review of Hot Deck Imputation for Survey Non-response." *International Statistical Review*, 78(1): 40–64.

Black, Dan A., Seth Sanders, and Lowell Taylor. 2003. "Causes and Consequences of Mismeasurement of Education in the Census." Syracuse University Working Paper.

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., and Barry T. Hirsch. 2013. "Is Earnings Nonresponse Ignorable?" *The Review of Economics and Statistics*, 95(2): 407–416.

Bonhomme, Stephane, and Jean-Marc Robin. 2010. "Generalized Non-parametric Deconvolution with an Application to Earnings Dynamics." *Review of Economic Studies*, 77(2): 491–533.

Brownstone, David, and Robert G. Valletta. 1996. "Modeling Earnings Measurement Error: A Multiple Imputation Approach." *The Review of Economics and Statistics*, 78(4): 705–717.

Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective. Monographs on Statistics and Applied Probability*. 2nd ed., Boca Raton:Chapman & Hall/CRC.

Chen, Xiaohong. 2007. "Large Sample Sieve Estimation of Semi-Nonparametric Models." In *Handbook of Econometrics*. Vol. 6b, , ed. James J. Heckman and Edward Leamer, Chapter 76, 5549–5632. Amsterdam:Elsevier.

Chen, Xiaohong, Han Hong, and Elie Tamer. 2005. "Measurement Error Models with Auxiliary Data." *The Review of Economic Studies*, 72(2): 343–366.

David, Martin, Roderick J. A. Little, Michael E. Samuhel, and Robert K. Triest. 1986. "Alternative Methods for CPS Income Imputation." *Journal of the American Statistical Association*, 81(393): 29–41.

Durrant, Gabriele B., and Chris Skinner. 2006. "Using Data Augmentation to Correct for Non-Ignorable Non-Response When Surrogate Data Are Available: An Application to the Distribution of Hourly Pay." *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 169(3): 605–623.

Greenlees, John S., William S. Reece, and Kimberly D. Zieschang. 1982. "Imputation of Missing Values when the Probability of Response Depends on the Variable Being Imputed." *Journal of the American Statistical Association*, 77(378): 251–261.

Hausman, Jerry A. 1978. "Specification Tests in Econometrics." *Econometrica*, 46(6): 1251–1271.

Heitjan, Daniel F., and Donald B. Rubin. 1991. "Ignorability and Coarse Data." *The Annals of Statistics*, 19(4): 2244–2253.

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.

Horowitz, Joel L., and Wolfgang Härdle. 1994. "Testing a Parametric Model against a Semiparametric Alternative." *Econometric Theory*, 10(5): 821–848.

Hsiao, Cheng. 1989. "Consistent estimation for some nonlinear errors-in-variables models." *Journal of Econometrics*, 41(1): 159–185.

Hu, Yingyao, and Geert Ridder. 2012. "Estimation of nonlinear models with mismeasured regressors using marginal information." *Journal of Applied Econometrics*, 27(3): 347–385.

Hu, Yingyao, and Susanne M. Schennach. 2008. "Instrumental Variable Treatment of Nonclassical Measurement Error Models." *Econometrica*, 76(1): 195–216.

Ichimura, Hidehiko, and Elena Martinez-Sanchis. 2005. "Identification and Estimation of GMM Models by Combining Two Data Sets." University College London Working Paper.

Ichimura, Hidehiko, and Elena Martinez-Sanchis. 2009. "Estimation and Inference of Models with Incomplete Data by Combining Two Data Sets." Unpublished Manuscript.

Imbens, Guido W., and Tony Lancaster. 1994. "Combining Micro and Macro Data in Microeconometric Models." *The Review of Economic Studies*, 61(4): 655–680.

Lee, Lung-fei, and Jungsywan H. Sepanski. 1995. "Estimation of Linear and Nonlinear Errors-in-Variables Models Using Validation Data." *Journal of the American Statistical Association*, 90(429): 130–140.

Levitan, Mark, and Trudi Renwick. 2010. "Using the American Community Survey to Implement a National Academy of Sciences-Style Poverty Measure: A Comparison of Imputation Strategies." American Statistical Association, Social Statistics Section Unpublished Manuscript.

Lillard, Lee, James P. Smith, and Finis Welch. 1986. "What Do We Really Know about Wages? The Importance of Nonreporting and Census Imputation." *Journal of Political Economy*, 94(3): 489–506.

Li, Tong. 2002. "Robust and consistent estimation of nonlinear errors-in-variables models." *Journal of Econometrics*, 110(1): 1–26.

Little, Roderick J. A. 2012. "Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics." *Journal of Official Statistics*, 28(3): 309–334.

Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical analysis with missing data.* . 2nd ed., New York:John Wiley.

Little, Roderick J. A., and Sonya Vartivarian. 2005. "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology*, 31(2): 161–168.

Meyer, Bruce D., Robert Goerge, and Nikolas Mittag. 2013. "Errors in Survey Reporting and Imputation and Their Effects on Estimates of Food Stamp Program Participation." Unpublished Manuscript.

Murphy, Kevin M., and Robert H. Topel. 1985. "Estimation and Inference in Two-Step Econometric Models." *Journal of Business & Economic Statistics*, 3(4): 370–379.

Newey, Whitney K., and Daniel L. McFadden. 1994. "Large Sample Estimation and Hypothesis Testing." In *Handbook of Econometrics*. Vol. 4, , ed. Robert F. Engle and Daniel L. McFadden, Chapter 36, 2111–2245. Amsterdam:Elsevier.

Rao, Jon N. K., and Jun Shao. 1992. "Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation." *Biometrika*, 79(4): 811–822.

Rao, Jon N. K., and Randy R. Sitter. 1995. "Variance Estimation Under Two-Phase Sampling with Application to Imputation for Missing Data." *Biometrika*, 82(2): 453–460.

Ridder, Geert, and Robert Moffitt. 2007. "The Econometrics of Data Combination." In *Handbook of Econometrics*. Vol. 6B, , ed. James J. Heckman and Edward Leamer, Chapter 75, 5469–5547. Amsterdam:Elsevier.

Rothe, Christoph, and Dominik Wied. 2013. "Misspecification Testing in a Class of Conditional Distributional Models." *Journal of the American Statistical Association*, 108(501): 314–324.

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

Rubin, Donald B. 1996. "Multiple Imputation After 18+ Years." *Journal of the American Statistical Association*, 91(434): 473–489.

Schafer, Joseph L., and Nathaniel Schenker. 2000. "Inference with Imputed Conditional Means." *Journal of the American Statistical Association*, 95(449): 144–154.

Schennach, Susanne M. 2004. "Estimation of Nonlinear Models with Measurement Error." *Econometrica*, 72(1): 33–75.

Schennach, Susanne M. 2007. "Instrumental Variable Estimation of Nonlinear Errors-in-Variables Models." *Econometrica*, 75(1): 201–239.

Sepanski, Jungsywan H., and Raymond J. Carroll. 1993. "Semiparametric quasilikelihood and variance function estimation in measurement error models." *Journal of Econometrics*, 58(1-2): 223–256.

Shao, Jun, and Randy R. Sitter. 1996. "Bootstrap for Imputed Survey Data." *Journal of the American Statistical Association*, 91(435): 1278–1288.

Smirnov, Nikolai V. 1939. "On the Estimation of the Discrepancy Between Empirical Curves of Distribution for Two Independent Samples." *Bulletin Mathematique de l'Universite de Moscou*, 2: 3–14.

Steuerle-Schofield, Lynne, Brian Junker, Lowell J. Taylor, and Dan A. Black. 2012. "Limitations of Institutional Plausible Values." Unpublished Manuscript.

Tarozzi, Alessandro, and Angus Deaton. 2009. "Using Census and Survey Data to Estimate Poverty and Inequality for Small Areas." *Review of Economics and Statistics*, 91(4): 773–792.

U.S. Census Bureau. 2009. "American Community Survey: Design and methodology." U.S. Census Bureau Technical Paper ACS-DM1, Washington, D.C.

Wooldridge, Jeffrey M. 2007. "Inverse probability weighted estimation for general missing data problems." *Journal of Econometrics*, 141(2): 1281–1301.