Bollinger, Christopher R.; Hirsch, Barry T.

Working Paper

Match bias from earnings imputation in the current population survey: the case of imperfect matching

IZA Discussion Papers, No. 1846

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Bollinger, Christopher R.; Hirsch, Barry T. (2005) : Match bias from earnings imputation in the current population survey: the case of imperfect matching, IZA Discussion Papers, No. 1846, Institute for the Study of Labor (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/33317

DISCUSSION PAPER SERIES

IZA DP No. 1846

Match Bias from Earnings Imputation in the Current Population Survey: The Case of Imperfect Matching

Christopher R. Bollinger
Barry T. Hirsch

November 2005

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Match Bias from Earnings Imputation in the Current Population Survey: The Case of Imperfect Matching

Christopher R. Bollinger
University of Kentucky

Barry T. Hirsch
Trinity University
and IZA Bonn

Discussion Paper No. 1846
November 2005

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
Email: iza@iza.org

IZA Discussion Paper No. 1846
November 2005

ABSTRACT

Match Bias from Earnings Imputation in the Current Population Survey: The Case of Imperfect Matching*

This paper examines alternative forms of match bias arising from earnings imputation. Wage equation parameters are estimated based on mixed samples of workers who do and do not report earnings, the latter group being assigned earnings of donors who share some but not all the attributes of the recipients. Regressions that include attributes not used as imputation match criteria (e.g., union status) are severely biased. Related forms of match bias arise with respect to attributes used as match criteria, but matched imperfectly. For example, an imperfect match on schooling creates bias that flattens estimated earnings profiles within low, middle, and high education groups, while creating large jumps in returns across groups. The same pattern arises in wage-age profiles. The paper provides a general analytic expression to correct match bias in regression coefficients under the assumption of conditional mean missing at random. The full sample correction approach is compared to the alternative of omitting imputed earners from the sample, with and without reweighting. Additional problems considered are bias in longitudinal analysis and the presence of dated donors.

JEL Classification: J31, C81, C10

Keywords: match bias, imputation, CPS, wage equations, measurement error

Corresponding author:

Barry T. Hirsch
Department of Economics
Trinity University
San Antonio, TX 78212-7200
USA
Email: bhirsch@trinity.edu

* Revised version of a paper presented at "Program Evaluation, Human Capital, and Labor Market Public Policy: A Research Conference in Honor of Mark C. Berger," October 7-8, 2004, Lexington, Kentucky. Previous versions have been presented at the Midwest Econometrics Group, Society of Labor Economics, and Western Economic Association Meetings, and at seminars at Florida State, Georgia State, Kentucky, and Ohio State. In addition to constructive suggestions from the editors and referees, helpful comments were received from John Abowd, Dan Black, David Blanchflower, J.S. Butler, Shiferaw Gurma, James Heckman, David Macpherson, Cordelia Reimers, MaryBeth Walker, and Aaron Yelowitz.

1 Introduction

In household surveys conducted by the Bureau of the Census, nonresponse rates for most questions are low. The striking exception is the high rate of nonresponse for questions on earnings and other sources of income. The chief reason for nonresponse is concern about confidentiality, although other reasons, such as insufficient knowledge among surveyed household members, matter as well (Groves and Couper 1998; Groves 2001). The approach most frequently employed by empirical researchers is to use imputed values provided by the Census. The implications of using the imputations provided by Census and others in estimation, however, are not well understood.[1] Lillard et al. (1986) warned that nonresponse and imputations in the March CPS significantly impacted conclusions about income and earnings. Recent work in the statistics literature (e.g., Wu 2004; Schafer and Schenker 2000) has focused upon inference with imputed values. Other work (Manski and Horowitz 1998, 2002) has focused on identification conditions when data are missing, but does not directly address the issue of using imputations. Hirsch and Schumacher (2004), whose work we extend, show that coefficient bias resulting from imputation of a dependent variable (earnings) can be of first order importance.

The Current Population Survey (CPS) monthly earnings files have earnings and wages imputed by the Census using a "cell hot deck" procedure in which Census "allocates" earnings to nonrespondents using an imputation method that assigns the reported earnings of a matched donor who has an identical mix of measured characteristics. The proportion of imputed earners was approximately 15% from 1979-1993, increased as a result of CPS revisions in 1994, and has risen in recent years to almost 30% (Hirsch and Schumacher 2004, Table 2). For a variety of reasons, the Census and the Bureau of Labor Statistics (BLS) choose to include both earnings respondents and nonrespondents in published tabulations of earnings and other outcomes of interest. Researchers typically do the same when estimating earnings equations, under the belief that including individuals with imputed earnings causes little bias in empirical results (Angrist and Krueger 1999, 1352-54). A recent paper by Hirsch and Schumacher (2004) shows that in a standard earnings equation, there exists attenuation or "match bias" toward zero for coefficients on those characteristics that are not imputation match criteria (e.g., union status). The attenuation is severe, roughly equal to the sample proportion with imputed earnings. Match bias (in effect, a bias resulting from non-match) operates independently of possible response bias, existing even when nonresponse is random (i.e.,

[1] For an excellent survey of imputation procedures, see Little and Rubin (2002), who state (p. 60): "Despite their popularity in practice, the literature on the theoretical properties of the various [hot deck] methods is very sparse."

missing at random).

Match bias associated with non-match attributes is a first-order problem. As shown in this paper, serious bias issues also arise with match attributes that are imperfectly matched. The Census uses broad categories to match donors' earnings with nonrespondents. For example, rather than matching on the exact age, individuals are grouped into six age categories. Similarly, Census uses three education categories – less than high school, high school through some college, and B.A. or above. When researchers include regressors in a wage equation containing greater detail than the match categories, say, detailed age or specific educational attainment levels, match bias can lead to highly misleading results.

This paper presents a general framework for examining match bias due to earnings imputation, deriving an analytic general bias measure under the assumption of conditional mean missing at random (CMMAR). Using this framework, we first formalize expressions for bias in the case of non-match dummy variables, the important case studied by Hirsch and Schumacher (2004). We then examine various cases of incomplete match. Even under the assumption of CMMAR, we show that biased wage regression estimates occur when including match attributes (e.g., schooling) at a level more detailed than used in the Census imputation match. We derive a set of corrections for incomplete match bias, demonstrate their use in several examples, and compare alternative approaches researchers might take to account for match bias.

The following section of the paper describes the Census cell hot deck imputation procedure. Subsequent sections examine alternative forms of match bias, in each case presenting analytic results showing the expected bias. Each section then demonstrates the issue at hand with empirical examples and compares alternative approaches to correct the bias. A final section provides informal analysis of other forms of bias arising from earnings imputation, the first involving bias in longitudinal estimates and the second resulting from matching the nominal earnings of "dated donors" to nonrespondents.

2 Census Earnings Imputation Methods in the CPS Monthly Earnings Files

Statistical agencies often impute or assign values to variables when an individual (or other unit of observation) does not provide a response or when a reported value cannot be shown because of confidentiality concerns. Imputation is common for earnings and other forms of income, where nonresponse rates are high. The appeal of imputation is that it allows data users to retain the

full sample of individuals which, with application of appropriate weights, blows up the sample into the full population. Often, imputation of one or a few variables makes it practical to retain an observation and use reported (non-imputed) information on other variables. Government agencies typically publish tables with descriptive data at relatively aggregate levels classified by broad categories (e.g., earnings by sex, age, and race). As long as the published classification categories are match criteria used in the imputation and not presented at a level narrower than in the imputation, inclusion of imputed earners does no harm. There is bias where presentation is for non-match criterion, say, earnings by union status and/or industry, or for classifications at finer levels, such as earnings by detailed rather than broad occupation.[2]

Analysis in this paper uses the CPS Outgoing Rotation Group (ORG) monthly earnings files, prepared by the Census for use by the Bureau of Labor Statistics (BLS), which then makes these files publicly available. An earnings supplement is administered to the quarter sample of employed wage and salary workers in their outgoing 4th and 8th months included in the survey. The sample design of the CPS is that individuals are included in the survey for eight months – four consecutive months in the survey, followed by eight months out, followed by four months in (the same months as in the previous year). The CPS-ORG earnings files begin in January 1979. They are typically used as annual files, including the twelve quarter samples during a calendar year.[3]

During 1979-93, approximately 15% of employed wage and salary workers have imputed values included for usual weekly earnings.[4] The CPS earnings questions were revised in 1994. The increased complexity and sequencing of earnings questions led to a substantial increase in imputation rates. Publicly available earnings files for January 1994 through August 1995 do not identify those with imputed earnings. Beginning September 1995, valid earnings allocation flags are included. Imputation rates have risen from about 22% in 1996 to about 30% in 2000-2004.

Earnings in the CPS-ORG are imputed using a "cell hot deck" method. There has been minor variation in the hot deck match criteria over time. For the ORG files during the 1979-1993 period, the Census created a hot deck or cells containing 11,232 possible combinations based on the following seven categories: gender (2 cells), age (6), race (2), education (3), occupation (13), hours worked

[2] BLS publishes an annual table compiled from the CPS earnings files that compounds these forms of bias, providing median weekly earnings for union and nonunion workers by industry and by occupation (the latter at a level more detailed than the imputation match). See U.S. Department of Labor (annual).

[3] Prior to 1979, the earnings supplement was administered to all rotation groups in May 1973 through May 1978. Nonrespondents are included in the May 1973-78 earnings files, but they do not have their earnings imputed. Approximately 20% of employed wage and salary workers in the May 1973-1978 files have no value (or the "missing" value) included in the usual weekly earnings field (Hirsch and Schumacher, 2004, Table 2).

[4] Earnings allocation flags are not reliable during 1989-93. Imputed earners can be identified based on those who do and do not have an entry in the "unedited" usual weekly earnings field (Hirsch and Schumacher, 2004).

(6), and receipt of tips, commissions or overtime (2). These categories are shown in Table 1. The Census keeps all cells "stocked" with a single donor, insuring that an exact match is always found. The donor in each cell is the most recent earnings respondent surveyed by the Census with that exact combination of characteristics. As each surveyed worker reports an earnings value, the Census goes to the appropriate cell, removes the previous donor value, and "refreshes" that cell with a new earnings value from the respondent.[5]

As shown in Table 1, the selection categories changed slightly in 1994 and 2003. Beginning in 1994, two additional hours cells were added for workers reporting variable hours, one for those who are usually full-time and one for those usually part-time, resulting in 14,976 possible combinations. Beginning in January 2003, the CPS adopted the 2000 Census occupation codes (COC), which involved a substantial revision from the 1980 and 1990 COC. Detailed occupation codes are grouped into 10 major categories, in contrast to 13 prior to 2003, resulting in 11,520 match cells.

At the start of each month's survey, cells are stocked with ending donors from the prior month. The Census retains donors until replaced, reaching back for donors as far as necessary, first within a given survey month and then to previous months and years. If needed, a donor value is used more than once. A donor's nominal earnings is assigned to the nonrespondent, with no adjustment for wage growth since the cell was refreshed. The Census does not retain information on cell refresh rates or the average "freshness" of donors. A trade-off exists. Less detailed match characteristics would produce more frequent refreshing of cells, but result in lower quality matches.[6]

[5] A brief discussion of Census/CPS hot deck methods is contained in the U.S. Department of Labor, 2002, p. 9.3). The more detailed information appearing here and in Hirsch and Schumacher (2004) was provided by economists at the BLS and Census Bureau. Unlike the ORGs, the March CPS Annual Demographic Files (ADF) use a "sequential" rather than "cell" hot deck imputation procedure to impute earnings (and income). Nonrespondents are matched to donors from within the same March survey in sequential steps, each step involving a less detailed match requirement. For example, suppose there were just four matching variables – sex, age, education, and occupation. The matching program would first attempt to find a match on the exact combination of variables using a relatively detailed breakdown. Absent a successful match at that level, matching proceeds to a next step with a less detailed breakdown, for example, broader occupation and age categories. Earnings imputation rates in the ADF are lower than in the ORGs. As emphasized by Lillard, Smith, and Welch (1986), the probability of a close match declines the less common an individual's characteristics. Although the imputation procedure used in the ADF produces a regression bias similar to that identified for the ORGs, our analysis applies most directly to the ORGs.

[6] Location is not an explicit match criterion. Files are sorted by location and nonrespondents are matched to the most recent matching donor. Thus, a donor is (roughly) the geographically closest person moving backward in the file. Nonrespondents with an unusually common mix of characteristics may be matched to someone in a similarly-priced neighborhood, creating a high quality match on some unmeasured as well as measured characteristics. More likely is that donors are found in different neighborhoods, cities, states, regions, or months. Once a match extends outside a current month, there should be no relationship between the locations of nonrespondents and donors. As seen subsequently, we estimate that 83% of nonrespondents are assigned the earnings of donors from previous survey months. In the March CPS, broad region serves as an explicit match criterion for selecting donors.

3 Imputation Match Bias

3.1 General Approach

In this section, we derive and report a general analytic approach to evaluate bias from the inclusion of imputed values in the dependent variable (much of this analysis is in the appendix). Following presentation of the general case, we examine specific cases of interest. We derive an analytic expression for bias in the case considered by Hirsch and Schumacher (2004), where an explanatory variable that is not an imputation match criterion is entered into a regression. We next consider two types of imperfect match. In the first case, a categorical variable such as educational degree or occupation is collapsed into broader categories for the purpose of imputation. In the second case, an ordinal variable which enters the regression, such as age, is collapsed into a set of categorical variables for the purpose of imputation. Finally, we consider a mixed case where a variable that enters the estimation equation as both a linear term and a categorical term, for example, years of education coupled with dummy variables for degree effect, is collapsed into broader categories.

Throughout this section, the variable y_i is the dependent variable in a linear regression, in this case the natural log of earnings. The variables $\underline{z}_i$ are the regressors of interest: age and education for example. The variables $\underline{x}_i$ represent the categories upon which matches are made. These variables are binary indicator (dummy) variables in practice, but our analysis does not rely upon this result. The full set of assumptions is presented in the appendix. We briefly review these assumptions below, with emphasis given to $A2$, conditional MAR.

$A1$: Only variable y_i is missing, for some but not all observations

$A2 : E_O\left[y_i|\underline{z}_i, \underline{x}_i\right] = E_M\left[y_i|\underline{z}_i, \underline{x}_i\right] = E\left[y_i|\underline{z}_i, \underline{x}_i\right]$

$A3 : \underline{x}_i = h\left(\underline{z}_i\right)$ where $h\left(\right)$ is a known deterministic function

$A4 : E\left[y_i|\underline{z}_i, \underline{x}_i\right] = E\left[y_i|\underline{z}_i\right] = \alpha + \underline{z}_i'\underline{\beta}$

$A5$: Imputed values of y_i are randomly drawn from the distribution $f_O\left(y_i|\underline{x}_i\right)$

Assumption $A1$ is self-explanatory. We examine the effect of measurement error in the dependent variable due to imputation, for some but not all observations. There is a large (and not unrelated) literature on right-hand side measurement error.

In assumption $A2$ and elsewhere, the notation $E_O\left[y_i|\underline{z}_i, \underline{x}_i\right]$ reads as the population expectation of y_i when y_i is Observed, while $E_M\left[y_i|\underline{z}_i, \underline{x}_i\right]$ is the population expectation of y_i for the Missing, those who do not report y_i and have earnings imputed. Assumption $A2$, which states that there is no selection on the y_i variable, is crucial. It assumes conditional missing at random, albeit in a

"weak" form, such that there are no difference in *mean* earnings between the observed and missing, conditional on $\underline{z}_i$. $A2$ allows the distribution of $(\underline{x}_i, \underline{z}_i)$ to differ between those who report earnings and those who don't. We say a "weak" form of MAR because it requires the mean, but not the distribution, of earnings within a match cell to be equivalent for those who report and do not report earnings. We refer to this as "conditional mean missing at random" or CMMAR. Although not formally considered here, $A2$ can be further weakened by allowing an intercept difference. Other research (Molinari 2005) considers cases where variables are not missing at random.

Assumption $A3$ is innocuous, simply stating that knowing $\underline{z}_i$ gives perfect information about the value of $\underline{x}_i$. That is, if you know the value of a variable at its detailed level, you know its value at an aggregated level. The opposite may not be true. Either $h()$ is many to one, as in the schooling and age cases, so $\underline{x}_i$ is a crude measure of $\underline{z}_i$, or there may be variables in $\underline{z}_i$ which are not measured in $\underline{x}_i$; for example, non-match characteristics union status, foreign born, and industry. An important implication for this is that $E[\underline{x}_i|z_i] = \underline{x}_i$, while $E[\underline{z}_i|\underline{x}_i]$ is not specified generally.

Assumption $A4$ implies that the relationship between y_i and $\underline{z}_i$ is linear in the parameters and that $\underline{x}_i$ do not contain information about y_i beyond what is contained in the more detailed variables $\underline{z}_i$. When $\underline{z}_i$ is categorical to begin with, this is always true, while when $\underline{z}_i$ is an ordinal variable, it implies that the specification is linear and there are no further nonlinearities that are better captured by the collapsed categories. Note that nonlinearities are allowed, the vector $\underline{z}_i$ must simply contain appropriate variables such as quadratic terms. Essentially, the assumption implies that the researcher has the correct specification for the conditional expectation function $E[y_i|\underline{z}_i]$.

Finally, assumption $A5$ implies that conditional upon $\underline{x}_i$, the distribution of the imputed y_i is independent of the distribution of $\underline{z}_i$. That is, the imputed data conditioned on $\underline{x}_i$ are independent of the variables not included as imputation match criteria.

We consider the population least squares projection of y_i on $\underline{z}_i$ when imputed values are used for those who do not report y_i. Under general assumptions, OLS is consistent for the least squares projection. The appendix formally derives the following important result for the estimated least squares slope coefficients b on variables $\underline{z}_i$:

$$b = \beta - p\left(E\left[\underline{z}_i\underline{z}_i'\right] - E[\underline{z}_i]\,E[\underline{z}_i]'\right)^{-1}\left(E_M\left[\underline{z}_i\left(\underline{z}_i - E_O[\underline{z}_i|\underline{x}_i]\right)'\right] - E[\underline{z}_i]\,E_M\left[\underline{z}_i - E_O[\underline{z}_i|\underline{x}_i]'\right]\right)\beta.$$

We refer to the expression above as the "general correction" for match bias, applicable in all cases discussed in this paper. The parameter p is the probability of not observing y_i (estimated by the proportion of missing values in the sample). Terms like $E_O[\underline{z}_i|\underline{x}_i]$ are the expectation of z_i given

x_i for the population who report y_i, while E_M is for the population who do not report y_i. Terms with no subscript are for the full population including both respondents and nonrespondents. The terms to the right of the initial β produce the match bias resulting from imputation.

The term $\left(E_M\left[\underline{z}_i\left(\underline{z}_i - E_O\left[\underline{z}_i|\underline{x}_i\right]\right)'\right] - E\left[\underline{z}_i\right]E_M\left[\underline{z}_i - E_O\left[\underline{z}_i|\underline{x}_i\right]'\right]\right)$ is the covariance between the regressors $\underline{z}_i$ and the prediction error from the relationship between those regressors and the match variables. Hence the entire term can be thought of in the following way: first regress z_i on the match variables and take the residuals $(z_i - E_O\left[\underline{z}_i|\underline{x}_i\right])$. Then regress those residuals back on z_i. This measures the variation in z_i that is not accounted for by the match variables. In essence this is measuring the omitted information from the imputation procedure and behaves like an omitted variable bias term. This can also be viewed as measurement error. The donor's earnings were generated from a particular value of z which does not necessarily match the value of z_i of the recipient. The measurement error is $(z_i - E_O\left[\underline{z}_i|\underline{x}_i\right])$, which measures the difference between the recipients z_i (the mismeasured variable) and the average donor's z_i for donors in the cell. The bias term is similar to the usual attenuation term found with measurement error.

Two simple cases illuminate the nature of match bias. First note that if $\underline{z}_i = \underline{x}_i$, implying that all variables in the model are included as imputation characteristics and at the same level of detail, then $b = \beta$ and no bias exists. Another interesting special case is where we have strict missing at random and z_i and x_i are scalars. In that case $E_M\left[\underline{z}_i - E_O\left[\underline{z}_i|\underline{x}_i\right]'\right] = 0$ and $E_M\left[\underline{z}_i\left(\underline{z}_i - E_O\left[\underline{z}_i|\underline{x}_i\right]\right)'\right]$ is the variance of z_i not explained by x_i. So, the ratio $E_M\left[\underline{z}_i\left(\underline{z}_i - E_O\left[\underline{z}_i|\underline{x}_i\right]\right)\right]/E\left[\underline{z}_i\underline{z}_i\right] - E\left[\underline{z}_i\right]E\left[\underline{z}_i\right] = 1 - V\left(z_i|x_i\right)/V\left(z_i\right)$, which is similar in concept to $1 - R^2$, but allows for a fully non-linear model. Indeed, in a case where x_i is binary (as is often the case for imputation characteristics), this is the R^2 from the regression of z_i on x_i. In the extreme case where $R^2 = 1$, all information in z_i can be accounted for by the imputation match criteria x_i, so there is no bias.

3.2 Standard Errors

Up to this point we have said nothing about bias in coefficient standard errors owing to imputation. Statistical significance is often not an issue in wage equation analysis owing to large sample sizes. Imputation does bias standard errors, however. Typical estimators of standard errors assume that observations are independent. When imputed values are drawn from other observations included in the sample, that assumption is violated. In general this will cause typical estimated standard errors to understate the true sampling variation. Heckman and LaFontaine (2004) examine the issue of standard errors in regressions using imputed values. Little and Rubin (2002) summarize

classic work addressing this issue.

Since the imputed observations are not independent of the non-imputed observations, the usual standard errors are not appropriate. Indeed, if the regression is y_i on $\underline{x}_i$, if all imputations are drawn from the observed sample, the standard errors reduce to the standard errors from only the observed sample. In the CPS hot deck procedure, some imputations derive from observations from previous months, some of which may not be included in the estimation sample. If the sample is selected on some $\underline{z}_i$ criteria (including time period), some imputations will be drawn from outside the criteria. In cases where the regression includes variables other than $\underline{x}_i$, as in the case studied here, there is some informational gain to including imputations.

Although one approach to estimating standard errors in this case would be to use a bootstrap, we use estimates based upon standard asymptotic results. Heteroskedastic robust standard errors for the OLS estimates are produced with typical software. To arrive at standard errors for the bias corrected results we assume non-stochastic regressors. The variance covariance matrix for the bias corrected slopes is then simply $A * V(b) * A^T$ where A is the bias correction matrix (since the estimates are simply Ab). This may tend to slightly understate the variance since it ignores variation in A. As in most empirical studies, we ignore the issue of sampling variation due to the imputations (Little and Rubin 2002).

In the following sections, we focus on specific forms of match bias, each permitting a simplification from the general case. Following theory presented in each section, we provide illustrative empirical evidence and apply the general bias correction developed here.

3.3 Non-Match Bias: Theory

Here we reconsider the results of Hirsch and Schumacher (2004), who examine the case of coefficient bias on a single non-match explanatory variables (e.g., union status). They present a bias expression for both a simple case where no other covariates are present in the regression and a general case where all other covariates are assumed to be exact match criteria.[7] The second case is an approximation based upon the results of Card (1996). We show that the approximation in Hirsch and Schumacher is quite close to the exact analytic result in most cases, but may differ substantially if a match characteristic is highly correlated with the non-match variable.

[7] In the simple case of no covariates, Hirsch and Schumacher (2004) show that the bias (the sum of the match error rates for union and nonunion nonrespondents) is equivalent to that for right-hand-side measurement error of a dummy variables, as developed by Aigner (1973) and extended in subsequent literature (e.g., Bollinger 1996; Black et al. 2000).

Let $\underline{z}_i = \begin{bmatrix} \underline{z}_{1i} \\ z_{2i} \end{bmatrix}$, where $\underline{z}_{1i} = \underline{x}_i$ and z_{2i} is a binary variable such as union status. All other covariates are included in the match criteria for imputation, but z_{2i} is not. Let $q = E[z_{2i}] = P(z_{2i})$, $q_M = E_M[z_{2i}]$, $q_O = E_O[z_{2i}]$, $q_M(\underline{z}_{1i}) = P_M[z_{2i}|\underline{z}_{1i}]$, $q_O(\underline{z}_{1i}) = P_O[z_{2i}|\underline{z}_{1i}]$, $V_{11} = V(\underline{z}_{1i})$, and $C = Cov(\underline{z}_{1i}, z_{2i})$, while R^2 is from the linear regression of z_{2i} on $\underline{z}_{1i}$ in the full population. Then the results in the appendix demonstrate that the LS coefficient on z_{2i} will be

$$\begin{aligned} b_2 &= \beta_2 \left(1 - p\left(\left(\frac{q_M - E_M[q_M(\underline{z}_{1i})\, q_O(\underline{z}_{1i})] - q\,(q_M - E_M[q_O[\underline{z}_{1i}]])}{(q-q^2)(1-R^2)}\right)\right.\right. \\ &\quad \left.\left. - \left(\frac{C'V_{11}^{-1}\left(E_M[\underline{z}_{1i}(q_M(\underline{z}_{1i}) - q_O(\underline{z}_{1i})] - E[\underline{z}_{1i}]\,(q_M - E_M[q_O(\underline{z}_{1i})])\right)}{(q-q^2)(1-R^2)}\right)\right)\right). \end{aligned}$$

The results of Hirsch and Schumacher (2004) provide an expression closely related to this, but based upon the assumption that the probability of misclassification is independent of the match criteria. This is an assumption of the results derived by Card (1996), which in turn were applied by Hirsch and Schumacher. If the strong missing at random assumption is applied, the two expressions are both equal to $\beta_1(1-p)$. Similarly, if $\underline{z}_{1i}$ and z_{2i} are uncorrelated the results are equivalent. The Hirsch and Schumacher results also do not extend to the case of multiple non-match variables. For these reasons, the general match bias correction derived in this paper is preferable.

3.4 Non-Match Bias: Evidence

In this section, we compare alternative methods to correct match bias, providing evidence on wage gap estimates with respect to selected attributes that are not match criteria. These gap estimates include union status, marital status, foreign born, Hispanic, and Asian, as well as wage dispersion across region, city size, and employment sectors (industry, public sector, and nonprofit status).[8] The sample is drawn from the CPS-ORG for 1998-2002. These years provide a convenient time period. Beginning in 1998, added information on education, including the GED, was included. Beginning in 2003, new occupation codes (from the 2000 Census) led to a change in the imputation match categories (see Table 1). Our estimation sample includes all non-student wage and salary employees ages 18 and over. Estimates are provided separately by gender, the sample of men being 388,578 and of women 369,762. In the male sample, 28.7% have earnings imputed, as compared to 26.8% of the female sample.

Table 2 provides coefficient estimates obtained from a standard log wage equation estimated using alternative approaches. Included in the equations are potential experience in quartic form

[8] Non-match attributes include not only variables measured in the monthly CPS, but also attributes measured in CPS supplements such as job tenure, employer size, and computer use.

(defined as the minimum of age minus years schooling minus 6 or years since age 16) and dummy variables for education (23 dummies), marital status (2), race/ethnicity (4), foreign-born, union, metropolitan size (6), region (8), occupation (12), employment sector (17), and year (4). The dependent variable is the natural log of average hourly earnings, including tips, commissions, and overtime, calculated as usual weekly earnings divided by usual weekly hours worked. Top-coded earnings are assigned the estimated mean above the cap ($2,885) based on an assumed Pareto distribution above the median (estimates are gender- and year-specific and roughly 1.5 times the cap, with small increases by year and higher means for men than for women).[9]

Wage gap estimates in Table 2 are drawn from regressions based on the full sample with Census imputations (the standard approach among researchers), the imputed ("missing") sample, the respondent ("observed") sample, the observed sample using inverse probability weighting (IPW) to correct for changes in the sample composition, and the full sample using the general bias correction derived in section 3.1. The IPW estimates require a brief explanation. Although we have assumed no specification error, in practice coefficients may differ across workers with different characteristics. If individuals are missing at random, the composition of the observed and full samples will be the same. If nonresponse is not random, estimates can differ. To account for the change in sample composition, we first run a probit equation with response as the binary dependent variable and all $\underline{z}_i$ as regressors. We then weight the observed sample by the inverse of the probability of response, thus giving enhanced weight to those most likely to be underrepresented in the observed sample (Wooldridge, 2002, pp. 587-588).

Severe match bias is readily evident in the estimates shown in Table 2. Focusing first on the male sample, the union-nonunion log wage gap is estimated to be .191 among respondents, only .024 among imputed earners, and .142 in the combined sample, a 25% attenuation (1-[.142/.191]), seen by the "Ratio (1)/(3)" column. Similar imputation bias is found for other non-match criteria. A married coefficient measures the wage gap between married males with spouse present and never married males. The full CPS sample produces an uncorrected marriage premium estimate of .096, while exclusion of imputed earners increases the estimate to .127, implying attenuation of 24%. The wage disadvantage for foreign-born workers is an estimated -.130 in the respondent sample, but only -.099 in the full sample. Hispanic workers have an estimated -.123 wage disadvantage using the respondent sample, compared to -.099 in the full sample. Wage gap estimates for Asian workers

[9] Mean earnings above the CPS cap by gender and year (since 1973), calculated by Barry Hirsch and David Macpherson, are posted at www.unionstats.com.

(compared to non-Hispanic whites) are small, but display similarly large attenuation (26%).

There exists a large literature on industry wage dispersion. Whatever one's interpretation of this literature, failure to account for match bias causes industry differentials (using wage level analysis) to be understated, since employment sector is not a Census match criterion. Table 2 provides wage dispersion estimates among 18 sectors, 13 private for-profit industry groups, 4 public sector groups (federal nonpostal, postal, state, and local), and the private nonprofit sector. The mean absolute log deviation for these 18 sectors is an estimated .117 based on the respondent sample, but falls to .090 using the full sample. One observes similar attenuation among wage differences for region and city size, standard control variables in most earnings equations.

Turning to the sample of women, we see exactly the same qualitative pattern seen among men. Magnitudes of the "worker attribute" wage gaps are somewhat smaller for women than for men. Interestingly, sectoral, region, and city size gaps are slightly larger among women. Attenuation from match bias is generally a little lower among women than men owing to a lower rate of nonresponse.

How do estimates based on the unweighted respondent sample compare to alternatives? Hirsch and Schumacher (2004) suggest that estimation from a respondent-only sample provides a reasonable first-order approximation of a true parameter, but may not fully account for match bias. In Table 2, we examine two alternatives to use of an unweighted respondent sample. Focusing on the union wage gap, we obtain a corrected full-sample union gap for men of .199, compared to a .191 based on the unweighted respondent sample; corresponding estimates for women are .148 and .143, respectively. These qualitative differences comport well with the results in Hirsch and Schumacher (2004).[10] If the differences between the corrected full samples and unweighted respondent samples is a result of composition differences, an attractive alternative may be to use a respondent sample weighted by the inverse of the probability of being in the respondent sample. These IPW results, shown in Table 2, produce a union wage gap estimate of .193 among men, higher than those obtained from the unweighted respondent sample but less than the corrected full sample estimate. The IPW union gap estimate is .143 among women, the same as the unweighted respondent estimate.

The patterns found for the union gap appear to be typical. As seen in Table 2, in all but one case the corrected full sample estimates exceed (in absolute value) estimates from the respondent

[10] When Hirsch and Schumacher (2004) estimate union wage gaps with the full sample, using either their own imputation procedure or correcting bias based on Card's measure for misclassification error, they obtain larger estimates than those obtained from the respondent sample. They suggest that attributes more common among nonrespondents are associated with larger union gaps. They do not explore whether the union result is common among a broader family of wage gap estimates.

sample (the exception is regional wage dispersion among men). The reweighted respondent sample (IPW) results among men tend to lie between the unweighted respondent and corrected full sample estimates. Among women, the IPW results are highly similar to the unweighted respondent results.

In Table 2, we present at the bottom of each ratio column significance tests for differences in all coefficients jointly across samples. For the male sample we obtain Wald statistics (ordered from high to low) of 991.2 for the uncorrected full versus corrected full, 285.3 for uncorrected full versus unweighted respondent, 101.7 for the uncorrected full versus weighted respondent, 39.5 for the unweighted respondent versus weighted respondent, 13.5 for the unweighted respondent versus corrected full, and 7.0 for the weighted respondent versus corrected full. Although all differences are significant (the critical value is 1.3), the magnitude of the difference between the corrected full sample and weighted respondent sample are relatively small. An identical qualitative pattern is found for women. Table 2 also summarizes results from significance tests (at the .05 level) for differences across regressions in coefficients for the five worker attribute non-match characteristics included in Table 2. In most cases the null of equality is readily rejected. Estimates are most similar among the corrected full and weighted respondent regressions (the far right column). Based on this comparison, we reject the null for "only" 2 of 5 coefficients among men and 3 of 5 among women.

Which estimation approach is preferred? This question is not easily answered. If we have the correct specification and conditional mean missing at random, as assumed in our bias correction, then the unweighted respondent sample, the weighted respondent sample (IPW), and the full sample with bias correction should produce consistent estimates. The only "wrong" approach is the standard one, including the full sample with Census imputations and no match bias correction. Difference between the corrected estimates from the full sample and those from the weighted and unweighted respondent samples result either from a violation of CMMAR or differences across groups in the value of the parameter of interest (i.e., specification error). None of these approaches accounts for a violation of CMMAR.[11]

When there exists specification error, some estimation approaches may be preferred to others. Researchers routinely choose (for good and bad reasons) to rely on simple but misspecified models. If a researcher desires a parameter estimate "averaged" across a representative population, then

[11] It is possible to account for nonignorable selection bias given appropriate instrument(s), but this is not a topic addressed in the paper. Hirsch and Schumacher (2004) estimate a selection model in which nonresponse is identified using as an instrument a variable indicating whether CPS survey questions are being answered by the individual or by another household member.

use of either the full sample with bias correction or the reweighted respondent sample is preferable to the unweighted respondent sample. Although an important contribution of this paper is the derivation and use of the full sample bias correction approach, it faces limitations for more general use. First, it is not trivial to understand and program, making it an unattractive approach for some researchers. Second, the bias correction derived here is designed specifically for the cell hot deck imputation used in CPS ORG, although the set up and its application can be used more broadly. The weighted respondent sample (IPW) approach may be more general, working well regardless of a survey's imputation methods, which may be highly complex or unknown to the researcher.[12] For these reasons, estimates from a reweighted respondent sample may be the preferred approach in a majority of applications. All of the approaches address the first-order match bias inherent in using the full uncorrected sample, but only IPW provides an easy and broadly applicable method to reweight the respondent sample to be representative of the full sample.

An alternative that we also briefly considered is to conduct one's own imputation (or multiple imputation) procedure, an approach that can be useful when tailored to a particular question at hand. For example, Hirsch and Schumacher (2004) conduct a simple cell hot deck imputation that adds union status as a match criterion, while Heckman and LaFontaine (2004) add the GED as an imputation match variable. Unfortunately, imputation is not an attractive *general* approach. A hot deck imputation that eliminates (or sharply reduces) discrepancies between the information provided by the included regressors $\underline{z}_i$ and the more limited Census match criteria $\underline{x}_i$ comes at a large cost. Adding imputation match criteria to a hot deck procedure leads to many thin and highly dated (or empty) cells. We explore a simple alternative. We conduct a regression-based imputation for nonrespondents using the predicted value from the observed sample parameters, plus an error term. Not surprisingly, this approach produces estimates highly similar to the unweighted

[12] The bias correction derived in this paper can be applied to either the CPS ORG cell hot deck or to the March CPS Annual Demographic File (ADF) sequential hot deck. Its assumptions, however, are more severely violated in the ADF. The bias correction assumes that the draw for the imputation is from the same distribution as the rest of the sample. The imputation draws from the conditional distribution $f(y|X_1, X_2)$ where the X's are the specific match characteristics. With dated donors from prior months, this is not literally true in the ORGs since $f(y_t|X_1, X_2)$ may differ from $f(y_{t-1}|X_1, X_2)$, but it is not a bad approximation. With the March ADF the assumption is violated when we draw from $f(y|X_1)$, the second or subsequent step matching only on some characteristics (an X at a broader level of detail). For both the ORG and the ADF, the question can be thought of as how different $f(y|X_1)$ is from $f(y|X_1, X_2)$. In general, there is probably less of a problem with ORG (last month's distribution is highly similar to this month's) than than with the ADF (the earnings distribution of male, HS grads, who work in a "narrow" occupation may be quite different than the distribution of male, HS grads for a "broad" occupation). For the ADF the questions are how often does the ADF move to matching with broader classifications and how different are those distributions? Lillard, Smith, and Welch (1986) show that broad matches are frequent and often poor. Thus, our general full sample correction method is probably not as good applied to the ADF as to the ORG. Weighted (IPW) respondent estimation is likely to be the better (as well as simpler) choice for use with the March CPS.

respondent sample results. It fails to account for composition bias owing to the use of the observed-only parameters and the absence of the detailed interactions implicit in a cell hot deck.

This section has demonstrated that attenuation of coefficients attached to variables not used as imputation match criteria is a concern of first-order importance and has compared alternative approaches to address match bias. In subsequent sections, the estimation approaches applied above for non-match attributes are used to account for bias from various forms of imperfect matching.

3.5 Imperfect Match on Multiple Categories

3.5.1 Theory

This section examines a less obvious form of match bias – bias for attributes that are match criteria, but are matched imperfectly. Specifically, we consider categorical variables $\underline{x}_i$ matched at a level more aggregated than seen among the included $\underline{z}_i$ regressors. The example we emphasize is education, where nonrespondents with detailed schooling level are assigned earnings from donors within one of three broad education groups. The same logic applies to other match criteria.[13] We previously presented a general bias formulation for this and other cases of match bias. Discussion below illustrates with some simple cases the nature of the bias in estimating returns to schooling.

Here we assume that $\underline{z}_i$ is a vector of $k-1$ binary variables representing k mutually exclusive categories (for example, educational categories). We assume that $x_i = 1$ represents the "last" J^* categories of $\underline{z}_i$ while $x_i = 0$ represents the reference category and the remaining categories of $\underline{z}_i$. Formally we define

$$x_i = \sum_{j \geq J^*} z_{ji}$$

where z_{ji} is the j^{th} element of $\underline{z}_i$.

As shown in the appendix,

$$\begin{aligned} E_s\left[y_i|\underline{z}_i\right] &= \left(\alpha + p\sum_{j=1}^{J^*-1}\Pr\left[z_{ji}=1|x_i=0\right]\beta_j\right) \\ &+ \sum_{j=1}^{J^*-1} z_{ji}\left(1-p\right)\beta_j \\ &+ \sum_{j=J^*}^{k-1} z_{ji}\left(\left(1-p\right)\beta_j + p\sum_{l=J^*}^{k-1}\Pr\left[z_{li}=1|x_i=1\right]\beta_l\right). \end{aligned}$$

[13] Only two imputation match criteria have perfect matching (ignoring reporting or recording error), sex and the receipt of overtime, tips, or commissions. Note that some match variables are ordered (e.g., age and hours worked) whereas others are not (e.g., occupation and race).

Thus, in the regression of y_i on $\underline{z}_i$ the intercept will be α plus p times a weighted average of the $\beta's$ for the z_{ji} where $x_i = 0$. The coefficients on $\underline{z}_i$ when $x_i = 0$ will be $(1-p)\beta_j$ and are simply downwardly biased. Finally the coefficients on the z_{ij} where $x_i = 1$, will be $(1-p)\beta_j$ plus p times the weighted average of all the β_j for z_{ji} where $x_i = 1$.

Consider a very simple case where there are four categories ($k = 4$) represented by three indicator variables ($k-1 = 3$), but two of the categories are combined for the match procedure ($J^* = 2$), which results in a binary match variable x_i. In the regression of y_i on z_{1i}, z_{2i}, and z_{3i},the intercept will be $\alpha + p\beta_1$. The coefficient on z_{1i} will be simply $(1-p)\beta_1$. Since $\Pr[z_{2i} = 1|x_i = 1] + \Pr[z_{3i} = 1|x_i = 1] = 1$, the coefficient on z_{2i} will be

$$b_2 = \beta_2 + p(\beta_3 - \beta_2)\Pr[z_{3i} = 1|x_i = 1].$$

If $\beta_3 > \beta_2$ the coefficient b_2 will be biased upward, while if $\beta_3 < \beta_2$, the coefficient b_2 will be biased downward.

In the more general case, we note that $\sum_{l=J^*}^{k-1} \Pr[z_{li} = 1|x_i = 1]\beta_l$ is a weighted average of the $\beta's$ for the $x_i = 1$ group. If β_j is less than this average, then the estimated coefficient will be inflated, while if β_j is more than this average it will be attenuated.

Since these results generalize in a straightforward way, this indicates that regressions with a full set of education dummy variables will have estimated returns to schooling that are biased. It is not difficult to extend the model to include other match variables. It is important to note that when other perfectly matched regressors are included as control variables their coefficients will be biased as well if they are correlated with the mismatched variables. Results in the appendix provide a general expression for bias in the linear regression setting.

3.5.2 Evidence: Returns to Schooling

Beginning in 1992, the CPS substituted an educational degree question for their previous measure of completed years of schooling. In 1998, additional questions were added to the CPS on receipt of a GED and years spent in school for both non-degree and degree students. Based on this information, one can construct detailed schooling degree/years variables that include well over 25 categories. One can also distinguish between years of schooling and highest degree, a "mixed" case examined in section 3.7. The ORG hot-deck imputation used since 1979 includes schooling as a match criterion, but matches the earnings of donors to nonrespondents based on three broad categories of education, which we label "low" (less than a high school degree), "middle" (a high school degree, including a GED, through some college), and "high" (a B.A. degree or above).

Were schooling the only match criterion, the expected value of donor earnings matched to nonrespondents would be the average earnings among respondents within each broad schooling category. Donor earnings would increase across the three schooling groups, but not within. Because other match criteria, in particular broad occupation, are correlated with schooling and earnings, imputed earnings may increase modestly within schooling groups. The schooling match creates an interesting form of match bias, flattening estimated earnings-schooling profiles within the low, middle, and high education groups, and creating large jumps across groups.

Figures 1a and 1b provide separate estimates of schooling returns for respondents and imputed earners. Estimates are from male and female wage equations using the same 1998-2002 CPS samples seen in the previous section. Shown in the figures are log wage differentials for each schooling group relative to earnings respondents with no schooling (set at zero). Control variables are listed in the figure note. Variables that most clearly reflect post-market outcomes (occupation, industry, union status, etc.) are not included.[14]

The basic story seen in the figures is identical for women and men. The earnings of respondents (shown by "diamonds) increase fairly steadily with schooling level. In contrast, imputed earnings among nonrespondents ("squares") are essentially flat in the low education category and increase slightly within the middle and high education categories. Failure to account for match bias leads to a downward bias in estimates for those at high education levels within each group and an upward bias for those with low education levels within each group. It leads to upwardly biased "jumps" in earnings as one moves across categories, specifically the movement from high school dropout to GED and from an associates degree to a B.A.

The GED results warrant examination. Here, upward match bias may be severe because the GED is the lowest education level within the middle education match category. Based on the sample of earnings respondents, the earnings gain for a male GED recipient relative to men who stop at 12 years of high school without a degree is a modest .036.[15] The same differential for imputed earners is an incredible .241 log points, seen in Figure 1a as the large jump between the Sch_12 and GED "squares." A standard wage equation using an uncorrected full sample would find a misleadingly large .087 wage gain for the GED (not shown), more than double the .036 estimate found for respondents. Similarly, imputation bias distorts the observed wage advantage for regular

[14] We do not interpret schooling parameters, even those corrected for match bias, as causal effects. Among other things, the estimates do not account for ability bias or reporting error in education.

[15] The CPS provides information on years of schooling completed prior to receipt of the GED. We do not use that information here, but do use it in subsequent analysis of "sheepskin" effects.

high school graduates as compared to GED recipients. The standard biased estimate indicates a .042 GED wage disadvantage, substantially smaller than the .072 GED disadvantage found among those with observed earnings. Among the sample of imputed earners, little wage difference is found between those with GEDs and standard diplomas. The story seen for women is highly similar.[16]

Equally startling examples of bias from imperfect matching are seen among workers with professional degrees and Ph.D.s. Match bias in this case is downward, owing to these groups having the highest education levels within the "high" schooling category. Nonrespondents with professional and Ph.D. degrees are matched to donors within the "high" schooling group, most of whom have a B.A. as their terminal degree. Estimates from the respondent sample reveal a large .355 log point wage advantage among men with professional degrees as compared to men with B.A. degrees. Based on a standard full sample without correction, the wage advantage is .241, attenuation being 32%. The bias is similarly large for women, a professional/B.A. degree wage advantage of .444 log points among earnings respondents, versus .296 using the full sample, attenuation of 33%. A similar pattern of bias is readily evident for those with Ph.D. degrees.

In short, match bias due to incomplete matching on education flattens wage-schooling profiles within educational match categories, while steepening the jump in wages between categories. Depending on the specific level of schooling attainment being examined, bias can range from small to very large. In a subsequent section, we examine a mixed model with an ordinal schooling variable and categorical degree variables (sheepskin effects).

3.6 Imperfect Match on Ordinal Variables

3.6.1 Theory

Here we consider a simplified case where a scalar ordinal variable, such as age, enters a regression linearly, but is reduced to two categories for the purposes of the imputation match. We use the term ordinal, but analysis in this section applies equally well to ordered categorical variables and cardinal variables. Indeed, age (or experience) is typically treated as cardinal. The specific structure is

$$E\left[y_i|z_i\right] = \alpha + \beta z_i$$

[16] Heckman and LaFontaine (2004) provide a detailed analysis of the GED and imputation bias, including a critique of misleading results found in Clarke and Jaeger (2002). Using the post-1998 CPS, they show that the positive effect of the GED on earnings is small once one omits imputed earners or, alternatively, use the GED as an imputation match criterion. Based on additional analysis using the NLSY, which permits an accounting for ability bias, the authors conclude that the remaining effects of the GED seen in the CPS are unlikely to be causal.

and

$$x_i = \begin{cases} 1 & \text{if } z_i > z^* \\ 0 & \text{if } z_i \leq z^*. \end{cases}$$

Given this simple structure, it follows then that

$$E\left[y_i|x_i\right] = \alpha + \beta E\left[z_i|x_i = 0\right] + \beta\left(E\left[z_i|x_i = 1\right] - E\left[z_i|x_i = 0\right]\right)x_i.$$

Substitution gives

$$E_s\left[y_i|x_i, z_i\right] = \alpha + (1-p)\,\beta z_i + p\beta\left(E\left[z_i|x_i = 0\right] + \left(E\left[z_i|x_i = 1\right] - E\left[z_i|x_i = 0\right]\right)x_i\right).$$

Then the linear projection of y_i on z_i gives an intercept of :

$$a = \alpha - \beta\left(p\left(1 - R^2\right)\right)E\left[z_i\right]$$

and a slope coefficient of

$$b = \beta\left(1 - p\left(1 - R^2\right)\right),$$

where R^2 is the correlation between z_i and x_i. The slope coefficient is attenuated or flattened by the proportion p imputed, mitigated in part by correlation between the information in match variables x_i and the non-match elements of z_i. We find this result generalizes to multiple categories and to the case of quadratic age: the quadratic profile will be flattened relative to the true profile when the imputed values are included.

Maintaining the assumption of missing at random, these results can be extended to the case where additional match characteristics are included in the regression. As with the previous case, all coefficients are biased.

3.6.2 Evidence: Wage-Age Profiles

As seen above, match bias resulting from imperfect matching arises in estimates of earnings profiles with respect to age (or potential experience). In the CPS, nonrespondents are matched to the earnings of donors in six age categories, ages 15-17, 18-24, 25-34, 35-54, 55-64, and 65 and over (our analysis includes nonstudent workers, 18 and over). Thus, the slopes of profiles are flattened within age categories, with jumps in earnings across categories. A simple way to illustrate the bias is to estimate linear wage-age profiles within each of the age categories using the respondent and imputed samples. We use a specification with largely "pre-market" demographic and schooling variables, plus location and year controls. These results are shown in Table 3.

The most notable bias is for young workers, whose wage-age profiles are steep. Focusing first on men, annual wage growth among respondents is .041 during ages 18-24 and .028 during ages 25-34. Wage growth seen among those with imputed earnings is far lower, .006 during ages 18-24 and .004 for ages 25-34. Wage growth is low in the 35-54 age interval, .005 in the respondent sample versus close to zero in the imputed sample. In the two oldest age categories, inclusion of imputed earnings causes wage decline to be understated. Identical patterns are seen for women, although overall wage-age growth is lower than for men (we observe wage growth with respect to age and not accumulated work experience). Whereas female respondents display annual wage growth of .029 during ages 18-24 and .020 during ages 25-34, growth using the imputed sample is effectively zero during these periods.

A more general way to illustrate the bias is to include a full set of age dummies and estimate wage-age profiles for respondents and nonrespondents. These results are shown for men and women in Figures 3a and 3b. Imputed earners exhibit substantial flattening of wage-age profiles within each age category, the bias being most serious during ages 18-24 and 25-34 when wage growth is highest. In the imputed worker sample, large wage jumps are observed between ages 24-25, 34-35, and, going in the opposite direction, 64-65. There is no jump between ages 54 and 55, since the weighted means of assigned donor earnings are similar in the adjacent age intervals.

Does inclusion of imputed earners greatly distort coefficients on potential experience in a Mincerian wage equation? The short answer is "a little." The most typical wage equation includes potential experience as a quadratic.[17] In a male wage equation, respondents have a quadratic log wage profile of .039 and -.068 (to rescale coefficients, Exp^2 is divided by 100). Estimates for the imputed sample produce a flatter profile, .035 and -.057. Estimating the profile using the full sample without correction, coefficient estimates are .038 and -.065, a profile slightly flatter than the one observed for respondents. Uncorrected standard errors (not shown) are much higher when imputed earners are included. An identical qualitative pattern is seen for women.

In short, bias due to imperfect matching causes wage patterns within and across age-match categories to be meaningless among imputed earners. Failure to account for this form of match bias has a modest effect in most wage equation applications, but should not be ignored in analyses of earnings-experience (age) profiles, particularly those focusing on wage growth among young workers.

[17] Murphy and Welch (1990) and Lemieux (2005) make strong arguments for use of higher order terms (e.g., up to a quartic) in the Mincerian wage equation, as was done in the regressions shown in Tables 2 and 4 and Figure 1.

3.7 Mixed Case: Imperfect Matching with Ordinal and Multiple Category Variables

3.7.1 Theory

Education provides an important example of a mixed case. Some researchers have observed that in addition to a linear return to years of education there appear to be "sheepskin" effects which result in jump discontinuities in the earnings-education profile. We examine the implications of match bias for this type of regression specification. Let z_{1i} be a dummy variable and z_{2i} be an ordinal variable, with

$$z_{1i} = \begin{cases} 1 & \text{if } z_{2i} > z^* \\ 0 & \text{otherwise} \end{cases}.$$

We assume that

$$E\left[y_i|\underline{z}_i\right] = \alpha + \beta_1 z_{1i} + \beta_2 z_{2i}$$

and that $x_i = z_{1i}$. That is the single match characteristic is the dummy variable. We maintain the CMMAR assumption here. Following the appendix, and recognizing that $x_i = z_{1i}$, the bias terms for the two slope coefficients will be

$$\begin{bmatrix} V_1 & C \\ C & V_2 \end{bmatrix}^{-1} \begin{bmatrix} 0 & E\left[z_{1i}\left(z_{2i} - E\left[z_{2i}|z_{1i}\right]\right)\right] \\ 0 & E\left[z_{2i}\left(z_{2i} - E\left[z_{2i}|z_{1i}\right]\right)\right] \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix},$$

where V_1 is the variance of z_{1i}, V_2 is the variance of z_{2i} and C is the covariance between z_{1i} and z_{2i}. The term $E\left[z_{1i}\left(z_{2i} - E\left[z_{2i}|z_{1i}\right]\right)\right] = 0$, while the term $E\left[z_{2i}\left(z_{2i} - E\left[z_{2i}|z_{1i}\right]\right)\right]$ is the variance of z_{2i} conditional on z_{1i}. Define R^2 as the squared correlation between z_{1i} and z_{2i} and note that $E\left[z_{2i}\left(z_{2i} - E\left[z_{2i}|z_{1i}\right]\right)\right] = V_2\left(1 - R^2\right)$. Then the above bias equation can be written as

$$\begin{bmatrix} V_1 & C \\ C & V_2 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 \\ 0 & V_2\left(1 - R^2\right) \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}.$$

Evaluating leads to the following expressions for the bias from the least squares projection

$$b_1 = \beta_1 + p\frac{C}{V_1}\beta_2$$

$$b_2 = \beta_2\left(1 - p\right).$$

Here we see that the degree effect will be overstated (since, by definition of z_{1i} and z_{2i} the covariance will be positive), while the year or marginal effect will be understated. Indeed, if there is no degree effect (if $\beta_1 = 0$), the OLS estimate will still be positive while the marginal effect will be understated. It must be kept in mind that the presence of other variables (whether non-match, imperfect match, or perfect match criteria) will alter these results.

3.7.2 Evidence: Sheepskin Effects and Linearity

A common approach in estimating the returns to schooling is to assume linearity and include a single schooling variable measuring years of school completed. The schooling coefficient represents the percentage (log) wage gain associated with an additional year of schooling (see Mincer 1974; Willis 1986; and subsequent literature for assumptions necessary to interpret this as a rate of return). A related approach includes indicators for completed degrees, measuring separately the effect of the "sheepskin" on earnings. This approach can be informative (but not decisive) in determining the extent to which education increases human capital and the extent to which it provides some verifiable signal of innate human capital or motivation. In the extreme (and ignoring complicating factors), if education is exclusively human capital enhancement, then the coefficients on the degree completion indicators should approach zero and years of schooling should measure the full human capital effect. If education provides only a signaling mechanism, then the coefficient on years schooling should approach zero and only the degree effects should matter.[18]

Table 4 provides estimates of a model with these mixed education variables. The sample is restricted to the range of data over which we can clearly distinguish between years of schooling and degree. We omit the relatively few workers with less than 9 years of schooling or with professional and Ph.D. degrees for whom separate information on years schooling is not provided.[19] Estimates are provided using the full sample with Census imputations and no bias correction (the standard approach), the respondent ("observed") sample, the observed sample with inverse probability weighting (IPW), and the full sample using the general correction measure derived in 3.1.

School is the measure of years schooling completed. The full sample estimate for men suggests a rate of return of .036 (in log points) for a year of schooling, holding degree constant. The estimate on the observed sample is .042 absent weights, and .043 reweighted to adjust for a changed sample composition. The corrected full sample estimate is .046, a percentage point larger than the uncorrected estimate. Some of the degree indicators, absent correction, are very misleading. For example, the coefficient on H.S. degree in the full sample is .136. The estimates from the observed sample, the IPW observed sample, and the corrected full sample are much smaller at .097, .094, and .092, respectively. Similarly, the estimated effect of a GED (years constant) is overstated due

[18] If unmeasured ability differences lead degree recipients to acquire more human capital per year of schooling than do nonrecipients, estimates of degree effects will be positively biased.

[19] M.A. recipients designate their program as a 1, 2, or 3+ year program. Information on additional years schooling is provided for those with some college and no degree and for B.A. degree recipients with graduate course work but no degree. Those with some college but no post-secondary degree are coded as having received a regular high school diploma (information on the GED is provided only for those without education beyond high school).

to match bias. The full sample estimate places the value of a GED at .119, while the observed, IPW observed, and full corrected sample estimate are only .067, .067, and .068, respectively.[20]

The results for women follow a similar pattern. The full sample return estimate of .048 is less than estimates from the observed sample of .054, the reweighted observed sample of .056, and the corrected full sample of .062. The GED full sample estimate of .129 compares to estimates of .091, .093, and .082 from the unweighted observed, IPW observed, and corrected full samples. Estimates of the value of a high school degree are highly similar to those seen among men.

The results confirm that imperfect group matching using the Census imputation procedure biases rate of return estimates, trivially for some schooling groups but substantially for others. In a sheepskin model, the Census imputation tends to understate the returns to years of schooling, while generally overstating degree effects. Sheepskin effects are still evident, but less pronounced than seen with observed Census earnings or from estimates corrected for match bias.

4 Additional Imputation Issues: Longitudinal Analysis and Dated Donors

4.1 Longitudinal Analysis

An increasing number of researchers take advantage of the longitudinal nature of the CPS to account for worker heterogeneity (fixed effects). Serious, although sometimes subtle, forms of match bias arise using longitudinal analysis with Census imputation values. Fortunately, researchers using CPS panels often omit workers with earnings imputed in either year one or two, in contrast to standard wage level analyses where researchers seldom do so. The decision to remove imputed earners in longitudinal analyses appears to result from a reluctance to introduce such a high degree of noise into the dependent variable, typically the change in log wages. Omitting imputed earners takes a large toll on reported (as opposed to meaningful) sample sizes. Matched worker panels are not nearly so large as unmatched cross sections in the first place, and wage effects are identified based on a relatively small number of workers changing status in the attribute of interest. Add to this the omission of those with earnings imputed in *either* of two years (38% of the potential sample in the example below) and one finds sample sizes of CPS panels to be rather modest.

We first discuss (but do not analyze) the case of match bias arising from non-match criteria in a longitudinal setting. Wage imputation in one or both years introduces spurious correlation

[20] Note that this estimate accounts for the years of schooling completed by GED recipients (mostly 9-12 years). Prior estimates of a GED effect, drawn from Figure 1, did not include a separate years schooling variable and compared GED recipients to those with 12 years schooling but no degree.

between measured wage change and the change in a non-match attribute z_i (say, union status) not included among the match criteria $\underline{x}_i$. If earnings are imputed in both years 1 and 2, there will be little if any correlation between the true change in z_i and the change in imputed wages. For those whose earnings are imputed in year 1 or year 2 but not both, the direction and magnitude of bias for each individual depends on whether earnings are imputed in year 1 or 2 and whether z_i is $[0, 1]$ or $[1, 0]$. Averaged across all workers with earnings imputed in one or both years, there is likely to be little relationship between wage change and, say, union status change. This exact result is noted (without explanation) by Card (1996, fn. 22). Longitudinal bias also arises when there is imperfect matching, a topic we do not address here.[21]

Bias arises in a rather more subtle form even when there is no mismatch with respect to the characteristic under study (i.e., a common $\underline{z}_i$ and $\underline{x}_i$). As an example, evidence is presented on the part-time/full-time wage gap. Part-time status is defined here as working fewer than 35 usual hours per week on the primary job. Census matches on eight categories of hours worked. There is no mismatch between the part-time status of nonrespondents and donors. Part-time nonrespondents are assigned the earnings of part-time donors; full-time nonrespondents are assigned earnings of full-time donors. Hence, part-time wage gaps estimated from wage level equations do not suffer from the match bias seen previously for non-match or imperfect match criteria.

Longitudinal studies identify the part-time wage effect based on workers who switch part-time status, from full- to part-time or vice versa, thus accounting for worker fixed effects Among switchers with imputed earnings, however, one does not "net out" fixed effects since the year-pair of wages is no longer for the same individual. Wage gap estimate are biased toward wage level results that fail to account for worker heterogeneity. This bias is illustrated in a study by Hirsch (2005) that examines part-time wage gaps. Part-time workers typically have less prior work experience and accumulated job skills (differences not measured by potential experience). Hence, longitudinal analysis accounting for worker heterogeneity is found to sharply reduce estimates of part-time penalties in comparison to wage level analysis.

In Table 5, we show reported and unreported results from the Hirsch study. Based on wage level estimates for respondents only (line 1-A), part-time penalties are -.087 for women (who account for 2/3 of part-time workers) and -.191 for men. All estimates include a detailed set of control variables. Line 1-B shows the same estimates for those whose earnings are imputed in either the

[21] The literature on longitudinal estimates of union wage gaps (e.g., Freeman 1984; Card 1996) has focused on attenuation due to measurement error in union status change. In the presence of misclassified union status, earnings imputation can mitigate or exacerbate measurement error bias.

current or prior year. The wage gap estimates are similar to those from the respondent sample, just as expected since there is no mismatch on part-time status. In fact, the estimated gaps from the imputed sample are a bit larger than for respondents.

Lines 2-A and 2-B of Table 5 provide estimates from longitudinal wage change equations. Using just the respondent sample (line 2-A), the estimated wage changes for those switching part-time status, either from FT to PT or from PT to FT, are effectively zero (separately, Hirsch [2005] finds wage changes for part-time switchers who also change occupation and industry). Estimates from the imputed earners sample (line 2-B), however, display estimates that are biased toward the wage level results. For example, women switching from FT to PT jobs are estimated as having a -.070 wage loss; men switching are estimated as having a -.18 loss. In short, part-time switchers who report their wages in consecutive years exhibit virtually no wage change relative to stayers. Switchers who do not report their earnings wrongly appear to have realized large wage changes. Although there is no mismatch on part-time status, PT treatment effects are being identified not from a sample of PT switchers but from a donor sample dominated by stayers.

4.2 Dated Donors

Earnings nonrespondents are assigned the nominal earnings of the donor who is the most recent respondent with an identical mix of match attributes. During the 1994-2002 period, the Census match procedure included 14,976 cells or combinations of match characteristics. For match cells with a relatively uncommon mix, donor earnings may be relatively dated, biasing downward imputed earnings owing to nominal and real wage growth. Stated alternatively, the survey month can be considered a wage determinant in $\underline{z}_i$ that, for nonrespondents, is imperfectly mapped from $\underline{x}_i$.

How serious is the dated donor problem? The Census does not record the "shelf age" of donor earnings assigned to nonrespondents. To assess this issue, one must approximate Census hot deck methods and measure the datedness of donor earnings. Our analysis begins with all employed wage and salary workers, ages 18 and over, from the December 2002 CPS. That month's file contains 4,759 nonrespondents. Some of these individuals will be matched to donor earnings in the current month, while most will reach back to donors in previous months and years. Each nonrespondent in December 2002 is given a unique match number corresponding to the 14,976 possible combinations of match attributes. Likewise, potential donors (respondents) in 60 monthly CPS earnings files (December 2002 back to January 1998) are assigned attribute match numbers on the same basis. We first examine whether at least one donor match exists for each nonrespondent in December 2002.

Those not finding a donor are retained and a search for a donor in November 2002 is executed. This process continues back to January 1998. In order to increase the size and representativeness of the nonrespondent sample, we conduct the identical analysis for nonrespondents during January-November 2002. The total number of nonrespondents during 2002 is 55,902.[22]

The outcome of the donor match exercise is shown in Figure 5, where cumulative match rates are presented. In the initial month, just 17.3% of 2002 nonrespondents find a same-month donor.[23] Reaching back one month, an additional 16.8% are matched, followed by 11.5% and 8.3% reaching back 2 and 3 months. Within these first 4 survey months (the sample month plus three months back), over half (53.9%) of all nonrespondents have been assigned donor earnings. Those not finding matches have lower and lower match hazards or probabilities of finding a match in subsequent months. Even after 5 years, reaching back 59 months from month zero to January 1998, 2.85% of nonrespondents remain without an earnings assignment and are assigned donor earnings in excess of 5 years old. In Figure 3, we add the residual monthly match rate of 2.85% to the prior month labeled 60+.

Beginning in 2003, the number of occupation categories in the Census match algorithm was reduced from 13 to 10, reducing the number of hot deck cells from 14.976 to 11,520. In order to see how this affects donor datedness, we provide an analysis matching the 17,864 earnings nonrespondents in January-April 2004 to donors beginning in April 2004 and reaching back to January 2003 (the first month with the new occupation codes). We find little change in average donor datedness. Whereas 53.9% of the 2002 nonrespondents found donors during the current or three previous months; the corresponding number for the January-April 2004 nonrespondents is 53.1%. Reaching back 15 months, 84.0% of the 2002 nonrespondents found a match; the corresponding number for 2004 respondents is 83.4%. We conclude that donor datedness has not appreciably changed as a result of the revised occupational match categories beginning in 2003.

How serious is the problem of dated donor earnings? Combining information on average donor age with the rate of wage growth, one can estimate the downward bias in average earnings. To

[22] For ease of programming, nonrespondents during each month of 2002 are treated as if they were December nonrespondents. That is, for each 2002 nonrespondent, we first search for matching donors in December 2002 and then reach back in time as far as January 1998.

[23] In order to approximate the Census match rate in the initial month, the donor pool is constructed by taking a 50% random sample of all respondents in December 2002. The Census searches for donors only among those who reside prior to the nonrespondent in the file layout (arranged geographically). Thus, nonrespondents at the beginning of the December 2002 file are necessarily assigned donors by the Census from November 2002 or earlier, whereas nonrespondents at the end of the December file can be matched to the full month donor sample. We approximate this by using a half donor sample in the initial month (and full samples thereafter). If we instead search through all December 2002 respondents for potential donors, the match rate in the initial month increases by several percentage points and the next month rate falls, with quick convergence in subsequent months to the rates shown in Figure 3.

calculate mean donor age one must assume an average match date for the nonrespondents who have failed to find a match in the previous five years. For the 2002 sample of nonrespondents, we assume that the 2.85% not matched going back to January 1998 would on average find a match in 6 additional months. Using this assumption, the average age or datedness of all donor earnings is 8.6 months or nearly $\frac{3}{4}$ of a year, substantially larger than the median age of 3 (the current month and three back).[24] If nominal wage growth were, say, 3% annually, this would imply that the average earnings of donors are understated by 2.25%. With approximately 30% of the CPS sample being nonrespondents, the CPS understates average earnings by .675% ($\frac{3}{4}$ of a year times 3% annual wage growth times .30 proportion donors) or two-thirds of a percentage point. In 2004, average hourly earnings compiled from the CPS, including imputed earners, is $17.69, 2.85% higher than the 2003 average of $17.20. Multiplying by .0064 (.75 times 2.85% times .30), earnings are understated by $.11, with the true average wage closer to $17.80. This was a period of modest nominal wage growth; the bias increases proportionately with the growth rate.

Do dated donors affect wage gap estimates? To the extent that a "treatment" group of workers has more (less) dated donors than a comparison group, the treatment group wage gap will be understated (overstated). Comparison of the average datedness of donors across various groups of workers based on gender and race, however, suggests that differences are not sufficiently large to substantively affect wage gap estimates standard in the literature.

Most CPS nonrespondents are matched to the nominal earnings of donors from prior months rather than the current month, causing earnings to be understated. The resulting bias for most labor market studies, however, is modest and does not warrant serious concern. If nominal wage growth were to increase sharply in future years, this conclusion would warrant reconsideration.

5 Conclusion

Match bias arising from Census earnings imputation is frequently an issue of some consequence, but one not generally considered by labor economists. Given the assumption of conditional mean missing at random (CMMAR), this paper derives a general analytic solution that measures match bias in its multiple forms. Bias is of first-order concern in those studies estimating wage gaps with respect to an attribute that is not a Census match criterion (union status, foreign born, etc.). Estimates can be obtained from samples including only earnings respondents (weighted or unweighted)

[24] The estimate of an 8.6 month mean donor age is sensitive to the assumed average match date for those relatively few (2.85%) nonrespondents remaining unmatched.

or coefficient estimates from the full sample can be corrected for match bias. Attenuation of coefficients attaching to non-match attributes can be roughly approximated by the proportion of imputed earners, nearly 30% in recent CPS earnings surveys.

This paper has shown that earnings imputation warrants concern in situations where there is matching on an attribute, but the match is imperfect (e.g., education, age, occupation). Matching across a range of values creates a form of match bias that flattens estimated earnings profiles within match categories (say, low, middle, and high education or age), while creating jumps across categories. Such bias can be modest or severe, leading either to overstatement (e.g., returns to the GED) or understatement (e.g., returns to professional and doctoral degrees).

We also draw attention to rather subtle forms of match bias. Severe bias may arises in longitudinal analysis even where there is no apparent mismatch (e.g., part-time status). Because true "switchers" (say between part-time and full-time work) are matched to the earnings of non-switching donors, estimates fail to net out worker fixed effects. Coefficient estimates are biased toward those obtained in standard wage level analysis. Another form of bias arises from wage understatement among nonrespondents due to the datedness of donor earnings, most imputed earners being assigned nominal earning from donors surveyed in prior months. At current rates of wage growth, however, bias from dated donors is unlikely to be a serious concern.

For the applied researcher, the simplest approach to account for match bias is to omit imputed earners from wage equation (and other) analyses. Alternatively, one can retain the full sample and calculate corrected parameter estimates as shown in this paper. Under the assumption of CMMAR and absent specification error, either set of parameter estimates is consistent. In practice, these approaches differ a bit. If one is concerned about composition effects, but does not wish to implement the analytic match bias correction outlined in the paper, a simple alternative is inverse probability weighted (IPW) least squares estimation on the respondent sample.[25] IPW has the added advantage of greater generality, being appropriate with surveys whose imputation methods differ substantively from the Census cell hot deck.[26]

Discussion in this paper has examined the CPS ORG earnings files and the estimation of earnings equations. Similar issues arise with the March CPS ADF and other household surveys, although rates of nonresponse are generally lower than in the ORGs and imputation methods (where used)

[25] Even ignoring match bias, a case can be made to use WLS with Census weights when using the full CPS sample, given that the CPS is not fully representative (Polivka 2000; Helwig, Ilg, and Mason 2001). Because our results were affected litle by the use of Census weights, we have not followed that approach.

[26] As discussed in section 3.4, for some applications, it may be practical to retain the full sample and implement one's own hot deck imputation using as a match variable the particular characteristic of interest.

differ from the cell hot deck. Although our focus has been on earnings imputation, similar issues arise for other variables whose values are imputed and are used as outcome (dependent) variables in empirical work. Fortunately, nonresponse rates on non-income related variables tend to be small. And, finally, earnings (income) is often used as an *explanatory* variable. If the dependent variable is not a Census match criterion, there will exist attenuation in the earnings coefficient for precisely the same reason seen in our discussion of match bias.

Ultimately, the moral of this story is that earnings imputation must be given more serious consideration by researchers than in the past.[27] Match bias resulting from earnings imputation is sometimes large and shows up in surprising places. Applied researchers should add match bias to their already long checklist of issues to consider. The Census and BLS should be more forthcoming about the precise methods used to impute earnings (income). Where an earnings variable is used either as a dependent or a key independent variable, researchers should use a sample of earnings respondents (unweighted or reweighted) or provide corrected full sample coefficient estimates. Inclusion of imputed earners absent bias correction should not occur, absent a persuasive argument for doing so. Such arguments are not easy to make.

[27] Our focus is on how researchers might best deal with current (and past) Census survey and imputation methods. Given the severity of the match bias problem, attention ought to be given as well to possible changes in Census imputation methods.

References

Aigner, Dennis J. 1973. "Regression with a Binary Independent Variable Subject to Errors of Observation." *Journal of Econometrics* 1: 49-59.

Angrist, Joshua D. and Alan B. Krueger. 1999. "Empirical Strategies in Labor Economics." In *Handbook of Labor Economics*, Vol. 3A, edited by Orley C. Ashenfelter and David Card. Amsterdam: Elsevier.

Black, Dan A., Mark C. Berger, and Frank A. Scott. 2000. "Bounding Parameter Estimates with Non-Classical Measurement Error." *Journal of the American Statistical Association* 95 (September): 739-48.

Bollinger, Christopher R. 1996. "Bounding Mean Regressions when a Binary Regressor is Mismeasured." *Journal of Econometrics* 73 (August): 387-99.

Card, David. 1996. "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis." *Econometrica* 64 (July): 957-79.

Clarke, Melissa A. and David A. Jaeger. 2002. "Natives, the Foreign-Born and High School Equivalents: New Evidence on the Returns to the GED." IZA Discussion Paper No. 477 (April).

Freeman, Richard B. 1984. "Longitudinal Analyses of the Effects of Trade Unions." *Journal of Labor Economics* 2 (January): 1-26.

Groves, Robert M. 2001. *Survey Nonresponse*. New Jersey: Wiley-Interscience.

Groves, Robert M. and Mick P. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: John Wiley.

Heckman, James J. and Paul A. LaFontaine. 2004. "Adjusting for CPS Wage Imputation Bias in Estimating Returns to GED Certification." Mimeographed, University of Chicago, March 15.

Helwig, Ryan T., Randy E. Ilg, and Sandra L. Mason. 2001. "Expansion of the Current Population Survey Sample Effective July 2001." *Employment and Earnings* 48 (August): 3-7.

Hirsch, Barry T. 2005. "Why Do Part-Time Workers Earn Less? The Role of Worker and Job Skills." *Industrial and Labor Relations Review* 58 (July): 525-551.

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

Horowitz, Joel L. and Charles F. Manski. 1998. "Censoring of Outcomes and Regressors Due to Survey Non-response: Identification and Estimation Using Weights and Imputations." *Journal of Econometrics* 84 (May): 37-58.

Horowitz, Joel L. and Charles F. Manski. 2000. "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data." *Journal of the American Statistical Association* 95 (March): 77-84.

Lemieux, Thomas. 2005. "The 'Mincer Equation' Thirty Years after *Schooling Experience, and Earnings*." In *Jacob Mincer, A Pioneer of Modern Labor Economics*, edited S. Grossbard-Shechtman. Springer Verlag, forthcoming.

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 (June): 489-506.

Little, Roderick J.A. and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*. New Jersey: Wiley-Interscience.

Mincer, Jacob. 1974. *Schooling, Experience, and Earnings.* New York: Columbia University Press.

Molinari, Francesca. 2005. "Missing Treatments." Mimeographed, Cornell University, June.

Murphy, Kevin M. and Finis Welch. 1990. "Empirical Age-Earnings Profiles." *Journal of Labor Economics* 8 (April): 202-229.

Polivka, Anne E. 2000. "Using Earnings Data from the Monthly Current Population Survey." Bureau of Labor Statistics, Mimeographed, October.

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

U.S. Department of Labor, Bureau of Labor Statistics. Annual. "Median Weekly Earnings of Full-time Wage and Salary Workers by Union Affiliation, Occupation and Industry." http://www.bls.gov/cps/cpsaat43.pdf.

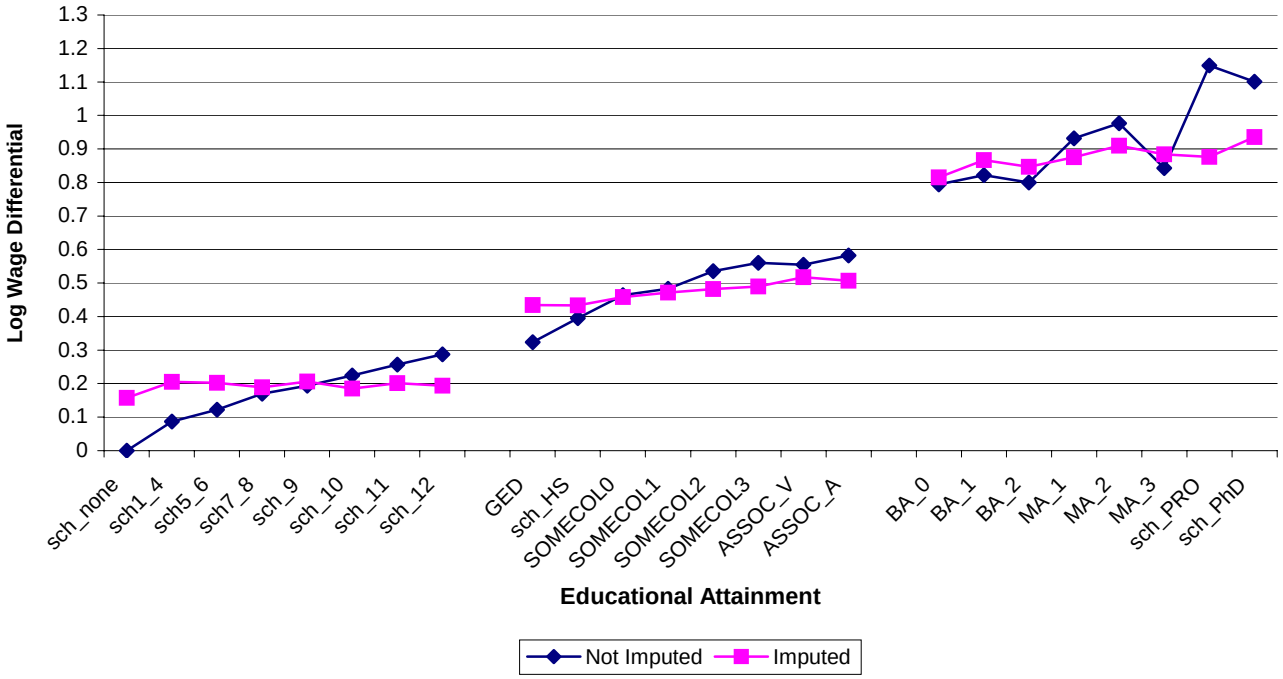
U.S. Department of Labor, Bureau of Labor Statistics. 2002. *Current Population Survey: Design and Methodology, Technical Paper 63RV* (March): www.bls.census.gov/cps/tp/tp63.htm.

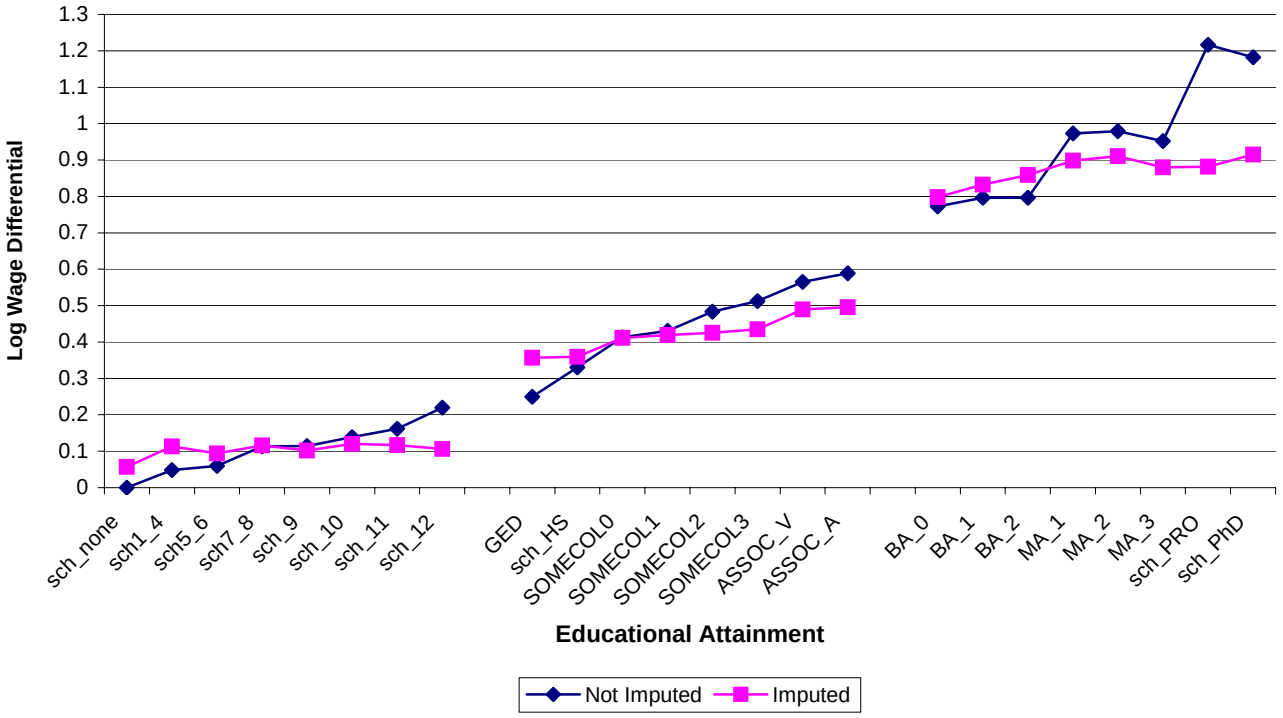
Willis, Robert J. 1986. "Wage Determinants: A Survey and Reinterpretation of Human Capital Earnings Functions." In *Handbook of Labor Economics*, Vol. 1, edited by Orley C. Ashenfelter and Richard Layard. Amsterdam: Elsevier.

Wu, Lang. 2004. "Exact and Approximate Inferences for Nonlinear Mixed Effects Models with Missing Covariates." *Journal of the American Statistical Association* 99 (September): 700-709.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Dection and Panel Data.* Cambridge: MIT Press.

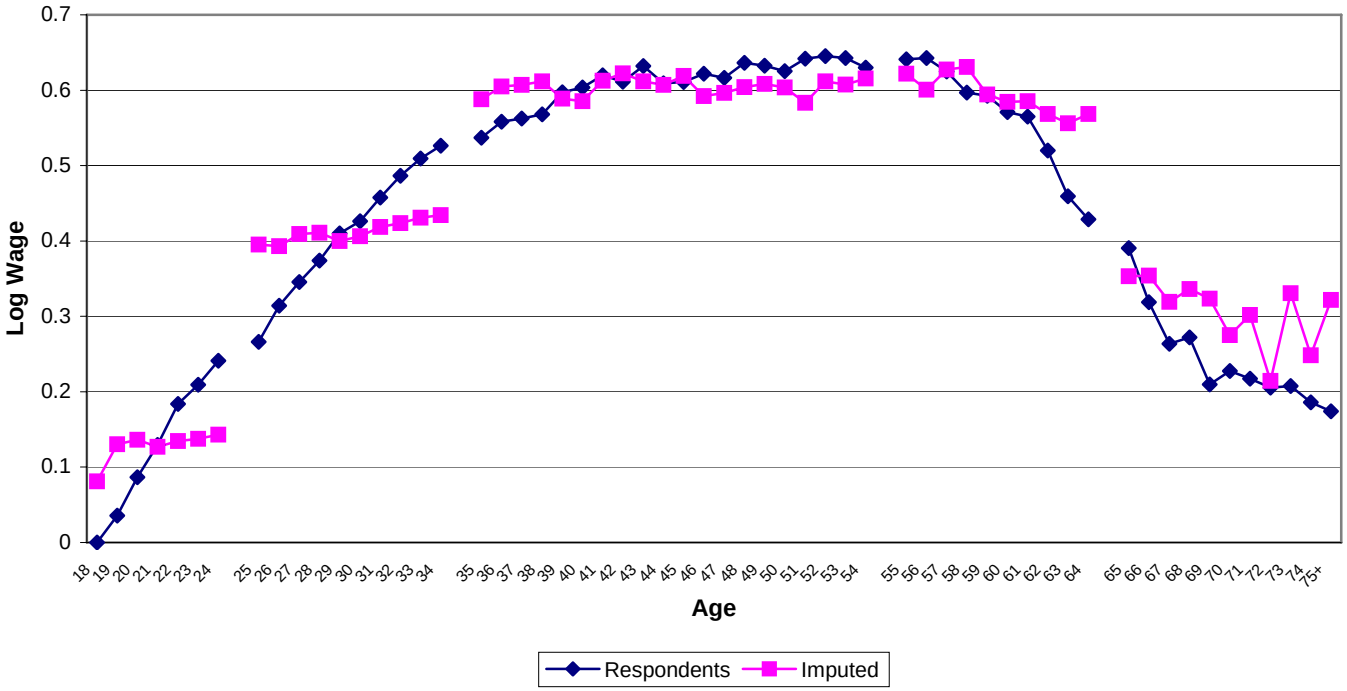
Figures 1a and 1b: Schooling Returns Among Male and Female Respondents and Imputed Earners, 1998-2002

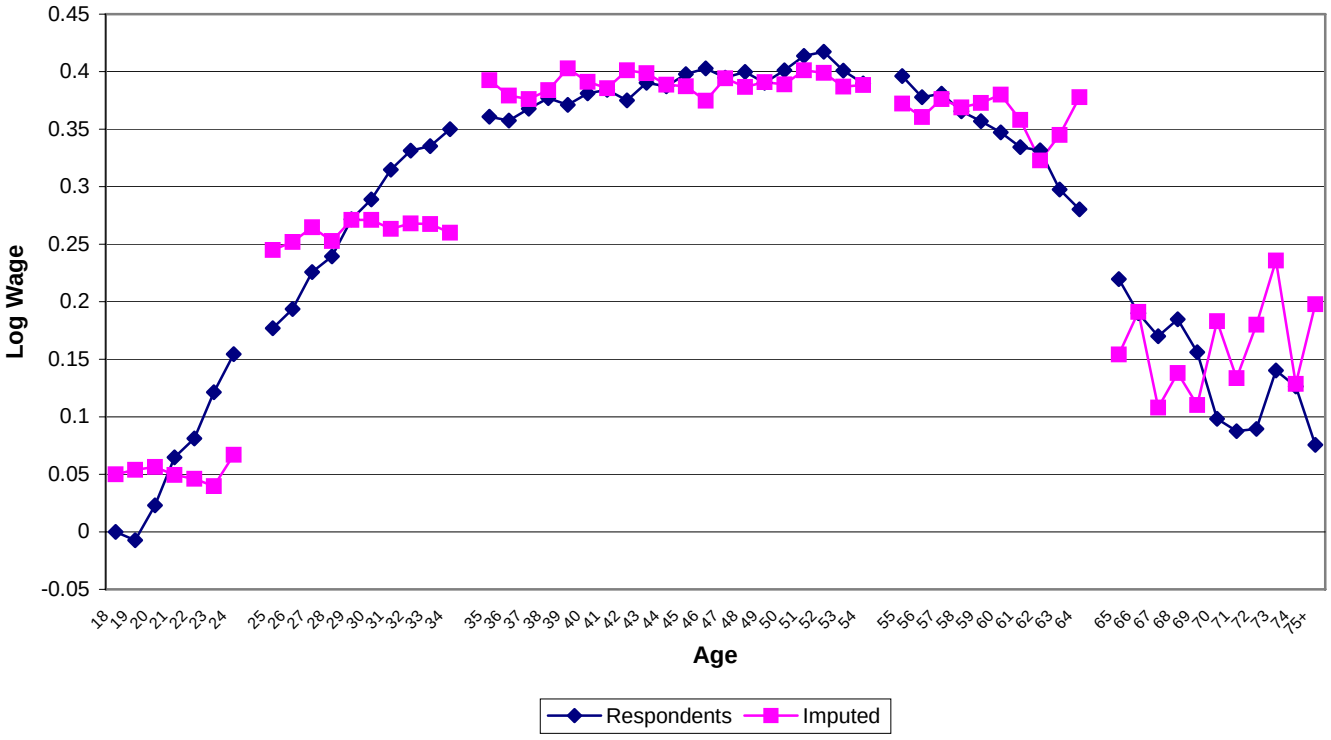




Estimates are from a pooled wage equation of respondents and imputed earners using the Current Population Survey monthly earnings files (CPS-ORG) for 1998-2002. The male sample size (top figure) is 388,578 – 276,909 respondents and 111,669 with earnings allocated (imputed) by the Census. The female sample size (bottom figure) is 369,762 – 270,537 respondents and 99,225 with earnings allocated (imputed) by the Census. The sample includes all non-student wage and salary workers, ages 18 and over. Shown are log wage differentials for each schooling group relative to earnings respondents with no schooling. In addition to the education variables, control variables include potential experience (defined as the minimum of age minus years schooling minus 6 or years since age 16) in quartic form, race-ethnicity (4 dummy variables for 5 categories), foreign-born, marital status (2), part-time, labor market size (6), region (8), and year (4).

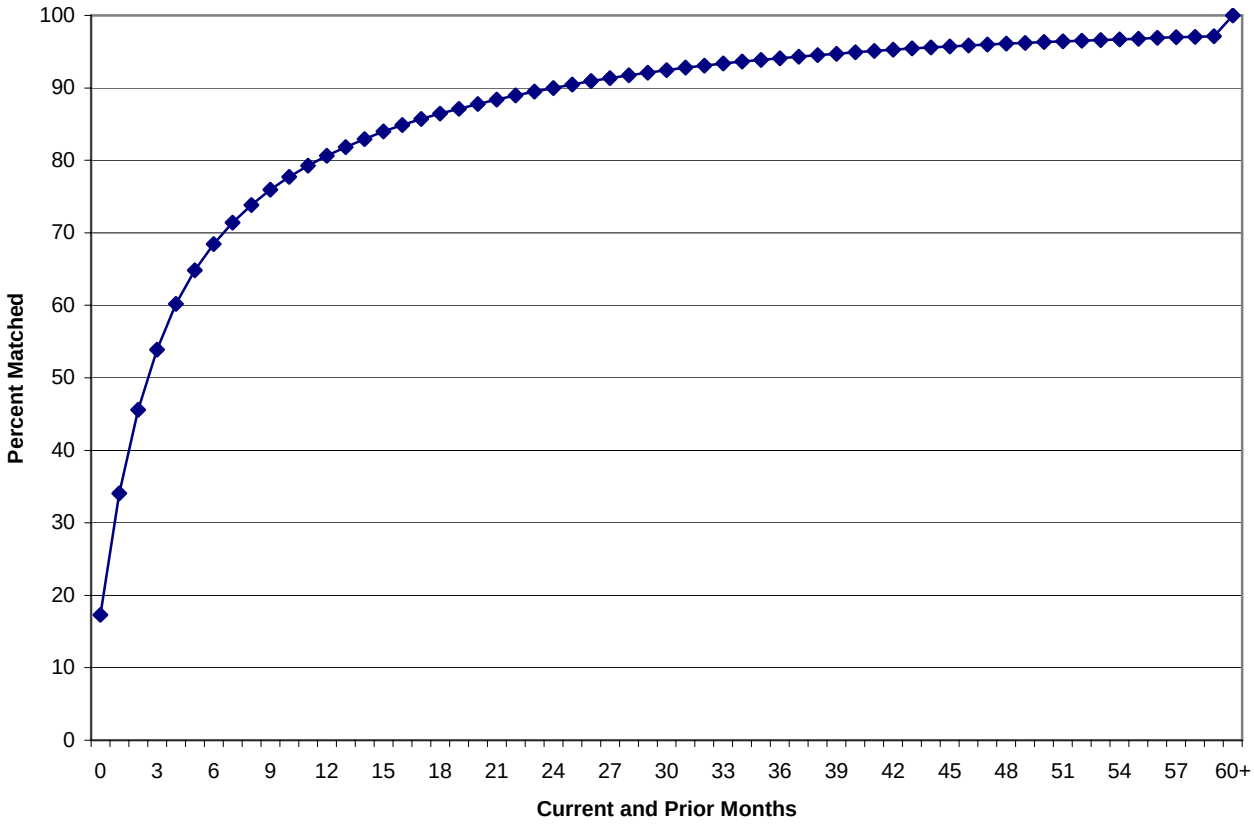
Figures 2a and 2b: Male and Female Wage-Age Profiles





Same samples as in Figures 1a and 1b. Shown are log wage differentials at each age relative to earnings respondents age 18. In addition to the education dummies, control variables include race-ethnicity (4 dummy variables for 5 categories), foreign-born, labor market size (6), region (8), and year (4).

Figure 3: Dated Donors: CPS Cumulative Imputation Match Rate for Current and Prior Month Donors



Cumulative monthly match rates of CPS-ORG nonrespondents in 2002 to 1998-2002 potential donors. Period 0 represents donor matches in the current survey month, while period n represents donor matches in the n^{th} prior month. Period 60+ figures represent all nonrespondents not finding a donor match during 1998-2002.

Table 1: CPS-ORG Cell Hot Deck Match Criteria, 1979 to Present

Match Criterion	Cells	Categories
1. Gender	2	Male / Female
2. Age	6	14-17 / 18-24 / 25-34 / 35-54 / 55-64 / 65+
3. Race	2	Black / Nonblack
4. Education	3	Less than high school High school through some college B.A. or above
5. Occupation (1979-2002)	13	Executive, administrative and managerial occupations Professional, specialty occupations Technicians and related support occupations Sales occupations Administrative support occupations, including clerical Private household occupations Protective service occupations Service occupations, except protective and household Precision production, craft and repair occupations Machine operators, assemblers and inspectors Transportation and material moving occupations Handlers, equipment cleaners, helpers and laborers Farming, forestry and fishing occupations
Occupation (2003-present)	10	Management, business, and financial occupations Professional and related occupations Service occupations Sales and related occupations Office and administrative support occupations Farming, fishing, and forestry occupations Construction and extraction occupations Installation, maintenance, and repair occupations Production occupations Transportation and material moving occupations
6. Hours Worked	8 (6)	0-20 / 21-34 / 35-39 / 40 /41-49 / 50+ Hours vary, usually full time (beginning 1994) Hours vary, usually part time (beginning 1994)
7. Overtime, Tips, or Commissions	2	Usually receive / Not usually receive
Total Imputation Cells:		1979-1993 11,232 1994-2002 14,976 2003-present 11,520

Source: Hirsch and Schumacher (2004) and information provided by Census and BLS economists. "Total imputation cells" is the product of the cell numbers shown. Beginning in 1994, designation for variable hours worked was introduced. Beginning in 2003, occupational categories were reduced from 13 to 10.

Table 2: Wage Gap Estimates Corrected and Uncorrected for Match Bias from Non-Match Criteria

	(1) Full Sample	(2) Imputed	(3) Respondents	(4) IP Weighted Respondents	(5) Corrected Full Sample	Ratio (1)/(3)	Ratio (1)/(4)	Ratio (1)/(5)	Ratio (3)/(4)	Ratio (3)/(5)	Ratio (4)/(5)
Men:											
Worker attribute coefficient:											
Union member	0.142	0.024	0.191	0.193	0.199	0.75*	0.74*	0.71*	0.99*	0.96*	0.97*
Married, spouse present	0.096	0.021	0.127	0.130	0.132	0.76*	0.74*	0.73*	0.97*	0.96*	0.99
Foreign born	-0.099	-0.024	-0.130	-0.133	-0.139	0.76*	0.75*	0.71*	0.98*	0.94*	0.96*
Hispanic	-0.099	-0.029	-0.123	-0.125	-0.128	0.81*	0.79*	0.77*	0.98*	0.96*	0.98
Asian	-0.024	-0.005	-0.033	-0.038	-0.038	0.74*	0.63*	0.63*	0.85*	0.86	1.00
Mean absolute deviation of coefficients:											
Sector-Ind/Pub/Nonprofit (18)	0.090	0.031	0.117	0.117	0.124	0.77	0.77	0.72	1.01	0.95	0.94
Metro size (7)	0.094	0.011	0.125	0.124	0.129	0.75	0.76	0.73	1.01	0.97	0.97
Region (9)	0.023	0.013	0.034	0.033	0.031	0.67	0.68	0.72	1.02	1.08	1.06
N / Wald statistic	388,578	111,669	276,909	276,909	388,578	285.3*	101.7*	991.2*	39.5*	13.5*	7.0*
Women:											
Worker attribute coefficient:											
Union member	0.111	0.013	0.143	0.143	0.148	0.78*	0.78*	0.75*	1.00	0.97*	0.97*
Married, spouse present	0.028	0.016	0.033	0.032	0.037	0.86*	0.87*	0.76*	1.01	0.88*	0.87*
Foreign born	-0.079	-0.015	-0.105	-0.103	-0.110	0.76*	0.77*	0.72*	1.01*	0.95*	0.94*
Hispanic	-0.077	-0.019	-0.096	-0.098	-0.100	0.80*	0.78*	0.77*	0.98*	0.96*	0.98
Asian	-0.016	0.002	-0.020	-0.023	-0.020	0.78	0.68*	0.78*	0.87*	0.99	1.14
Mean absolute deviation of coefficients:											
Sector-Ind/Pub/Nonprofit (18)	0.098	0.030	0.128	0.128	0.133	0.77	0.77	0.74	1.00	0.96	0.96
Metro size (7)	0.102	0.018	0.129	0.129	0.135	0.79	0.79	0.76	1.00	0.96	0.96
Region (9)	0.040	0.012	0.052	0.051	0.053	0.78	0.78	0.76	1.01	0.97	0.96
N / Wald statistic	369,762	99,225	270,537	270,537	369,762	200.5*	75.7*	681.5*	24.2*	18.1*	9.8*

The sample includes all non-student wage and salary workers ages 18 and over, from the January 1998-December 2002 monthly CPS-ORG earnings files. The proportion of the full CPS sample with imputed earners is .287 among men and .268 among women. Results are shown for the full sample (respondents plus nonrespondents with Census imputed earnings), imputed (missing) earners only, earnings respondents (observed) only, respondents with inverse probability weighting (IPW), and the full sample with parameter estimates corrected by the general match bias measure. Included in the wage equation are potential experience in quartic form and dummy variables for education (23 dummies), marital status (2), race/ethnicity (4), foreign-born, part-time, union, metropolitan size (6), region (8), occupation (12), employment sector (17), and year (4). Sector includes 18 groups: 13 private for-profit industry categories, private nonprofit, and the public sector groups postal, federal non-postal, state, and local. Shown in the top panel are log wage gaps with the following reference groups: union vs. nonunion workers, married with spouse present vs. single, foreign-born vs. U.S. born, Hispanic vs. non-Hispanic white, and Asian vs. non-Hispanic white. Shown in the bottom panel is the mean absolute deviation of coefficients (unweighted) with the omitted reference group counted as zero. The first three ratio columns show observed attenuation coefficients, the ratio of the uncorrected to alternative corrected estimates. The last three columns show the ratios of corrected estimates. The * shown next to the ratios indicate that the null of equal coefficients on the given variable between the designated columns can be rejected at the .05 significance level. The * shown next to the Wald statistics applies to the null of jointly equivalent coefficients between the designated equations.

Table 3: Wage-Age and Wage-Experience Profile Estimates

	Men	Women
1. Linear wage growth per year within age groups		
Respondents		
18-24	.041	.029
25-34	.028	.020
35-54	.005	.002
55-64	-.021	-.011
65 plus	-.013	-.010
Imputed earners		
18-24	.006	.001
25-34	.004	.002
35-54	.000	.000
55-64	-.007	-.002
65 plus	-.003	.004
2. Quadratic potential experience profiles		
Respondents		
Exp	.039	.025
$Exp^2/100$	-.068	-.044
Imputed earners		
Exp	.035	.023
$Exp^2/100$	-.057	-.039
Pooled sample		
Exp	.038	.024
$Exp^2/100$	-.065	-.042
Sample Sizes:		
Respondents	276,909	270,537
Imputed	111,669	99,225
Pooled	388,578	369,762

Source: CPS-ORG, 1998 – 2002; all non-student wage and salary workers, ages 18 and over. Control variables include a full set of education dummies, demographic variables, region, city size, and year. Specifications including age variables do not include potential experience.

Table 4: Estimated Schooling and Sheepskin Effects, 1998-2002

	Full Sample	Imputed	Respondents	IP Weighted Respondents	Corrected Full Sample
Men**:**					
School (years completed)	0.036	0.022	0.042	0.043	0.046
GED	0.119	0.251	0.067	0.067	0.068
High School	0.136	0.230	0.097	0.094	0.092
Associates Degree	0.190	0.270	0.156	0.151	0.160
B.A.	0.367	0.549	0.294	0.287	0.268
Masters	0.414	0.587	0.345	0.337	0.335
N	359,564	103,476	256,088	256,088	359,564
Women**:**					
School (years completed)	0.048	0.030	0.054	0.056	0.062
GED	0.129	0.236	0.091	0.093	0.082
High School	0.137	0.224	0.104	0.104	0.088
Associates Degree	0.237	0.290	0.215	0.213	0.214
B.A.	0.368	0.562	0.297	0.293	0.252
Masters	0.440	0.595	0.382	0.375	0.347
N	353,585	95,120	258,465	258,465	353,585

Source: CPS-ORG, 1998 – 2002; all non-student wage and salary workers, ages 18 and over with between 9 years schooling and a masters degree (omitted are those with schooling less than 9, professional degrees, and Ph.D.s). Control variables include a full set demographic variables, region, city size, and year. Full sample includes both the respondent (observed) and imputed (missing) samples with Census imputation. Corrected estimates are based on the full sample and the general bias correction shown in the text. The IP weighted column reports least squares estimates from the respondent sample reweighed by the inverse probability that an individual's earnings are reported.

Table 5: Effects of Imputation on Panel Fixed Effects: Part-time/Full-time Log Wage Gaps

	Men	Women
1. Standard Wage Level Equation		
A. Respondents		
Part time, current	-.191	-.087
B. Imputed Earners		
Part time, current	-.237	-.109
2. Wage Change Equations		
A. Respondents		
FT to PT	.006	.020
PT to FT	-.016	-.008
B. Imputed Earners		
FT to PT	-.184	-.070
PT to FT	.128	.048

Source: Hirsch (2005, Tables 2-3). CPS-ORG Panels, September 1995/96-2001/02. Wage level equations shown above are for the 2nd year. Detailed controls are included; identical qualitative results are obtained with no controls. The respondent only sample excludes workers with earnings imputed in either the current or prior year (38.2% of a pooled male sample and 37.0% of a pooled female sample). The respondent samples include 88,576 men and 88,161 women; the imputed samples include 54,713 men and 51,670 women.

1 Appendix

1.1 Notation and Assumptions

The work here is related to work developed by Horowitz and Manski (1998, 2002). Their work examined general identification results. This work specifically examines the implications of including imputations.

Let $f(y_i, \underline{x}_i, \underline{z}_i, R_i)$ be the population distribution of the variables of interest. The variable y_i is the dependent variable of interest. The vector $\underline{z}_i$ is a set of regressors of interest. The vector $\underline{x}_i$ is a set of variables upon which imputation will be based. The variable R_i is a binary indicator which equals O if y_i is observed, and equals M if y_i is missing. We define the distribution $f_O(y_i, \underline{x}_i, \underline{z}_i) \equiv f(y_i, \underline{x}_i, \underline{z}_i | R_i = O)$ for individuals who do report all variables. We define $p = \Pr[R_i = M]$.

$A1$: Only the variable y_i is missing for some observations.

We define $f_M(y_i, \underline{x}_i, \underline{z}_i) = f(y_i, \underline{x}_i, \underline{z}_i | R_i = M)$ for individuals who do not report y_i. The distribution $f_M(y_i, \underline{x}_i, \underline{z}_i)$ is not observed. However, $f_M(\underline{x}_i, \underline{z}_i)$ is observed. That is, only y_i is missing.

$A2$: Conditional Missing at Random: $f_O(y_i|\underline{x}_i, \underline{z}_i) = f_M(y_i|\underline{x}_i, \underline{z}_i) = f(y_i|\underline{x}_i, \underline{z}_i)$.

Conditional upon both $(\underline{x}_i, \underline{z}_i)$ the distribution of y_i is the same in both the observed and unobserved populations. We actually only require that $E_O[y_i|x_i, z_i] = E_M[y_i|x_i, z_i] = E[y_i|x_i, z_i]$, which is somewhat weaker.

$A3$: $\underline{x}_i = h(\underline{z}_i)$,where $h(.) : R^{J_1} \longrightarrow R^{J_2}$.

The function h is a known deterministic function. For some elements z_{ji}, the function h is the identity function. This represents the fact that some variables may be the same both in the specification and in the imputation match criteria, while other variables will be reduced to cruder categorical variables.

$A4$: $E[y_i|\underline{z}_i, \underline{x}_i] = \alpha + \underline{z}_i\beta = E[y_i|\underline{z}_i]$.

The $\underline{z}_i$ contain all relevant information, inclusion of $\underline{x}_i$ does not provide additional information about the mean of y_i. That is, the researcher has the correct specification.

$A5$: Imputed values of y_i are randomly drawn from the distribution $f_O(y_i|\underline{x}_i)$.

Conditional upon $\underline{x}_i$, the distribution of imputed y_i is independent of the distribution of $\underline{z}_i$. Hence the joint distribution of $(y_i, \underline{z}_i)$ condition upon the match criteria $\underline{x}_i$ is given by

$$f_I(y_i, \underline{z}_i|\underline{x}_i) = f_O(y_i|\underline{x}_i) * f_M(\underline{z}_i|\underline{x}_i) \tag{1}$$

Hence, the joint distribution of the imputed data can be written as

$$f_I(y_i, \underline{x}_i, \underline{z}_i) = f_O(y_i|\underline{x}_i) * f_M(\underline{z}_i|\underline{x}_i) * f_M(\underline{x}_i) = f_O(y_i|\underline{x}_i) * f_M(\underline{x}_i, \underline{z}_i) \tag{2}$$

This highlights the independence of the imputed data from variables which are not used in the imputation. The joint distribution from which the sample is drawn can be written as a mixture distribution:

$$\begin{aligned} f_s(y_i, \underline{x}_i, \underline{z}_i) &= (1-p) f_O(y_i, \underline{x}_i, \underline{z}_i) + p f_I(y_i, \underline{x}_i, \underline{z}_i) \\ &= (1-p) f_O(y_i, \underline{x}_i, \underline{z}_i) + p f_O(y_i|\underline{x}_i) * f_M(\underline{x}_i, \underline{z}_i). \end{aligned} \tag{3}$$

We term this distribution the Sample Distribution.

Throughout, $E_O[*] = \int *f_O$ and $E_M[*] = \int *f_M$. That is, expectations within the sub population where $R_i = O$ or $R_i = M$, respectively.

1.2 Conditional Expectation Function in Sample Distribution

The definition of the distributions above implies that

$$f_s(\underline{x}_i, \underline{z}_i) = f(\underline{x}_i, \underline{z}_i) = (1-p) f_O(\underline{x}_i, \underline{z}_i) + p f_M(\underline{x}_i, \underline{z}_i). \tag{4}$$

By definition

$$f_s(y_i|\underline{x}_i, \underline{z}_i) = \frac{f_s(y_i, \underline{x}_i, \underline{z}_i)}{f(\underline{x}_i, \underline{z}_i)}. \tag{5}$$

Substitution yields

$$f_s(y_i|\underline{x}_i, \underline{z}_i) = (1-p)\frac{f_O(y_i, \underline{x}_i, \underline{z}_i)}{f(\underline{x}_i, \underline{z}_i)} + p\frac{f_O(y_i|\underline{x}_i) * f_M(\underline{x}_i, \underline{z}_i)}{f(\underline{x}_i, \underline{z}_i)}. \tag{6}$$

Multiplying and dividing the first term in 6 by $f_O(\underline{x}_i, \underline{z}_i)$ and rearranging yields

$$f_s(y_i|\underline{x}_i, \underline{z}_i) = (1-p) f_O(y_i|\underline{x}_i, \underline{z}_i)\frac{f_O(\underline{x}_i, \underline{z}_i)}{f(\underline{x}_i, \underline{z}_i)} + p f_O(y_i|\underline{x}_i)\frac{f_M(\underline{x}_i, \underline{z}_i)}{f(\underline{x}_i, \underline{z}_i)}. \tag{7}$$

The conditional expectation function of y_i given $(\underline{x}_i, \underline{z}_i)$ for the Sample Distribution is then derived by integration

$$E_s[y_i|\underline{x}_i, \underline{z}_i] = \int y_i f_s(y_i|\underline{x}_i, \underline{z}_i)\, dy$$

$$= (1-p)\left(\int y f_O(y_i|\underline{x}_i, \underline{z}_i)\, dy\right)\frac{f_O(\underline{x}_i, \underline{z}_i)}{f(\underline{x}_i, \underline{z}_i)} + p\left(\int y f_O(y_i|\underline{x}_i)\, dy\right)\frac{f_M(\underline{x}_i, \underline{z}_i)}{f(\underline{x}_i, \underline{z}_i)}$$

$$= (1-p) E_O[y_i|\underline{x}_i, \underline{z}_i]\frac{f_O(\underline{x}_i, \underline{z}_i)}{f(\underline{x}_i, \underline{z}_i)} + p E_O[y_i|\underline{x}_i]\frac{f_M(\underline{x}_i, \underline{z}_i)}{f(\underline{x}_i, \underline{z}_i)}. \tag{8}$$

From assumption A4, this becomes

$$E_s[y_i|\underline{x}_i, \underline{z}_i] = (1-p)\left(\alpha + \underline{z}_i'\beta\right)\frac{f_O(\underline{x}_i, \underline{z}_i)}{f(\underline{x}_i, \underline{z}_i)} + p\left(\alpha + E_O[\underline{z}_i|\underline{x}_i]'\beta\right)\frac{f_M(\underline{x}_i, \underline{z}_i)}{f(\underline{x}_i, \underline{z}_i)}. \tag{9}$$

Adding subtracting $p\underline{z}_i\beta\frac{f_M(\underline{x}_i, \underline{z}_i)}{f(\underline{x}_i, \underline{z}_i)}$, noting that $\frac{(1-p)f_O(\underline{x}_i, \underline{z}_i)}{f(\underline{x}_i, \underline{z}_i)} + \frac{p f_M(\underline{x}_i, \underline{z}_i)}{f(\underline{x}_i, \underline{z}_i)} = 1$ and combining terms yields

$$E_s[y_i|\underline{x}_i, \underline{z}_i] = \alpha + \underline{z}_i'\beta - p(\underline{z}_i - E_O[\underline{z}_i|\underline{x}_i])'\beta\frac{f_M(\underline{x}_i, \underline{z}_i)}{f(\underline{x}_i, \underline{z}_i)}. \tag{10}$$

1.3 Least Squares Projection

Under general assumptions Ordinary Least Squares is consistent for the slope coefficients in the least squares projection of y_i on the variables of interest in the population. The least squares projection of y_i on $\underline{z}_i$ in the Sample Distribution is defined by

$$\arg\min_{a,b} E_s\left[\left(E_s\left[y_i|\underline{x}_i,\underline{z}_i\right]-a-\underline{z}_i'b\right)^2\right].$$

Using the definition of $E_s\left[y_i|\underline{x}_i,\underline{z}_i\right]$ in equation 10 yields

$$=\arg\min_{a,b}\int\left(\alpha+\underline{z}_i'\beta-p\left(\underline{z}_i-E_O\left[\underline{z}_i|\underline{x}_i\right]\right)'\beta\frac{f_M\left(\underline{x}_i,\underline{z}_i\right)}{f\left(\underline{x}_i,\underline{z}_i\right)}-a-\underline{z}_i'b\right)^2 f\left(\underline{x}_i,\underline{z}_i\right)dxdz. \tag{11}$$

The first order conditions to the minimization problem are give by

$$\int\left(\alpha+\underline{z}_i'\beta-p\left(\underline{z}_i-E_O\left[\underline{z}_i|\underline{x}_i\right]\right)'\beta\frac{f_M\left(\underline{x}_i,\underline{z}_i\right)}{f\left(\underline{x}_i,\underline{z}_i\right)}-a-\underline{z}_i'b\right)f\left(\underline{x}_i,\underline{z}_i\right)dxdz=0$$

and

$$\int\underline{z}_i\left(\alpha+\underline{z}_i'\beta-p\left(\underline{z}_i-E_O\left[\underline{z}_i|\underline{x}_i\right]\right)'\beta\frac{f_M\left(\underline{x}_i,\underline{z}_i\right)}{f\left(\underline{x}_i,\underline{z}_i\right)}-a-\underline{z}_i'b\right)f\left(\underline{x}_i,\underline{z}_i\right)dxdz=0. \tag{12}$$

Regrouping and simplifying yields

$$\alpha+E\left[\underline{z}_i\right]'\beta-pE_M\left[\underline{z}_i-E_O\left[\underline{z}_i|\underline{x}_i\right]\right]'\beta-a-E\left[\underline{z}_i\right]'b=0$$

and

$$E\left[\underline{z}_i\right]\alpha+E\left[\underline{z}_i\underline{z}_i'\right]\beta-pE_M\left[\underline{z}_i\left(\underline{z}_i-E_O\left[\underline{z}_i|\underline{x}_i\right]\right)'\right]\beta-E\left[\underline{z}_i\right]a-E\left[\underline{z}_i\underline{z}_i'\right]b=0. \tag{13}$$

Note that if $p=0$ (no imputations), then these are the "normal equations" and result in the usual $a=\alpha$ and $b=\beta$ result. Thus without imputations, the OLS estimator will be consistent for α,β.

Solving these yields

$$\begin{aligned}a &= \alpha+pE\left[\underline{z}_i\right]'\left(E\left[\underline{z}_i\underline{z}_i'\right]-E\left[\underline{z}_i\right]E\left[\underline{z}_i\right]'\right)^{-1}\left(E_M\left[\underline{z}_i\left(\underline{z}_i-E_O\left[\underline{z}_i|\underline{x}_i\right]\right)'\right]-E\left[\underline{z}_i\right]E_M\left[\underline{z}_i-E_O\left[\underline{z}_i|\underline{x}_i\right]'\right]\right)\beta\\ &\quad -pE_M\left[\underline{z}_i-E_O\left[\underline{z}_i|\underline{x}_i\right]\right]\beta\end{aligned}$$

and

$$b=\beta-p\left(E\left[\underline{z}_i\underline{z}_i'\right]-E\left[\underline{z}_i\right]E\left[\underline{z}_i\right]'\right)^{-1}\left(E_M\left[\underline{z}_i\left(\underline{z}_i-E_O\left[\underline{z}_i|\underline{x}_i\right]\right)'\right]-E\left[\underline{z}_i\right]E_M\left[\underline{z}_i-E_O\left[\underline{z}_i|\underline{x}_i\right]'\right]\right)\beta. \tag{14}$$

The OLS slope estimates can be corrected by the formula

$$\beta=\left(I-p\left(E\left[\underline{z}_i\underline{z}_i'\right]-E\left[\underline{z}_i\right]E\left[\underline{z}_i\right]'\right)^{-1}\left(E_M\left[\underline{z}_i\left(\underline{z}_i-E_O\left[\underline{z}_i|\underline{x}_i\right]\right)'\right]-E\left[\underline{z}_i\right]E_M\left[\underline{z}_i-E_O\left[\underline{z}_i|\underline{x}_i\right]'\right]\right)\right)^{-1}b,$$

the intercept can be obtained by rewriting the equation for a, or, more simply, by using the consistent estimates for β in the formula

$$E\left[y_i\right]-E\left[\underline{z}_i\right]\beta.$$

Provided that conditional missing at random holds, $E\left[y_i\right]$ can be estimated by either the sample average of the observed sample or the full sample including imputations. The term $E\left[\underline{z}_i\right]$ can be estimated by sample averages as well.

1.4 Binary Non-Match Variable

Here we derive the results for section 3.1, the case considered by Hirsch and Schumacher (2004). Let $\underline{z}_i = \begin{bmatrix} \underline{z}_{1i} \\ z_{2i} \end{bmatrix}$, where $\underline{z}_{1i} = \underline{x}_i$ and z_{2i} is a binary variable and partition the vectors β and b conformably. Note that $E_O[\underline{z}_i|\underline{x}_i] = \begin{bmatrix} \underline{z}_{1i} \\ E_O[z_{2i}|\underline{x}_i] \end{bmatrix}$ and so $\underline{z}_i - E_O[\underline{z}_i|\underline{x}_i] = \begin{bmatrix} 0 \\ z_{2i} - E_O[z_{2i}|\underline{x}_i] \end{bmatrix}$. Let $q = E[z_{2i}] = P(z_{2i})$, and let $q_M = E_M[z_{2i}]$, $q_O = E_O[z_{2i}]$, $q_M(\underline{z}_{1i}) = P_M[z_{2i}|\underline{z}_{i1}]$, $q_O(\underline{z}_{1i}) = P_O[z_{2i}|\underline{z}_{i1}]$, $V_{11} = V(\underline{z}_{i1})$, $V_{22} = V(z_{2i})$, $C = Cov(\underline{z}_{i1}, z_{i2})$ and R^2 is the r-squared from the linear regression of z_{i2} on $\underline{z}_{i1}$ in the full population. Equation 14 becomes

$$\begin{aligned} b &= \beta - p \begin{bmatrix} V_{11} & C \\ C' & V_{22} \end{bmatrix}^{-1} \left(\begin{pmatrix} 0 & E_M\left[\underline{z}_{1i}\left(z_{2i} - E_O[z_{2i}|\underline{z}_{1i}]\right)'\right] \\ 0 & E_M\left[z_{2i}\left(z_{2i} - E_O[z_{2i}|\underline{z}_{1i}]\right)'\right] \end{pmatrix} \right. \\ &\quad \left. - \begin{pmatrix} 0 & E[\underline{z}_{1i}] E_M[z_{2i} - E_O[z_{2i}|\underline{z}_{1i}]] \\ 0 & E[z_{2i}] E_M[z_{2i} - E_O[z_{2i}|\underline{z}_{1i}]] \end{pmatrix} \right) \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}. \end{aligned}$$

Noting that

$$\begin{bmatrix} V_{11} & C \\ C' & V_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \left(V_{11} - CV_{22}^{-1}C'\right)^{-1} & -V_{11}C'\left(V_{22} - C'V_{11}^{-1}C\right)^{-1} \\ -\left(V_{22} - C'V_{11}^{-1}C\right)^{-1} CV_{11} & \left(V_{22} - C'V_{11}^{-1}C\right)^{-1} \end{bmatrix},$$

$\left(V_{22} - C'V_{11}^{-1}C\right)^{-1} = V_{22}\left(1 - R^2\right)$, and $V_{22} = q - q^2$ since z_{2i} is binary, substitution and evaluation yield:

$$\begin{aligned} b_1 &= \beta_1 - p\beta_2 \left(\left(\left(V_{11} - \frac{CC'}{q - q^2} \right) \left(E_M[\underline{z}_{i1}(q_M(\underline{z}_{i1}) - q_O(\underline{z}_{i1})] - E[\underline{z}_{i1}]\left(q_M - E_M[q_O(\underline{z}_{1i})]\right) \right) \right) \right. \\ &\quad \left. - V_{11}^{-1}C' \left(\frac{q_M - E_M[q_M(\underline{z}_{i1}) q_O(\underline{z}_{i1})] - q\left(q_M - E_M[q_O[\underline{z}_{i1}]]\right)}{(q - q^2)(1 - R^2)} \right) \right) \end{aligned}$$

and

$$\begin{aligned} b_2 &= \beta_2 \left(1 - p \left(\left(\frac{q_M - E_M[q_M(\underline{z}_{i1}) q_O(\underline{z}_{i1})] - q\left(q_M - E_M[q_O[\underline{z}_{i1}]]\right)}{(q - q^2)(1 - R^2)} \right) \right. \right. \\ &\quad \left. \left. - \left(\frac{C'V_{11}^{-1}\left(E_M[\underline{z}_{i1}(q_M(\underline{z}_{i1}) - q_O(\underline{z}_{i1})] - E[\underline{z}_{i1}]\left(q_M - E_M[q_O(\underline{z}_{1i})]\right)\right)}{(q - q^2)(1 - R^2)} \right) \right) \right). \end{aligned}$$

1.5 Discussion of Bias Terms

Under Missing at Random, $E_M[\underline{z}_i - E_O[\underline{z}_i|\underline{x}_i]] = 0$ and $E_M\left[\underline{z}_i\left(\underline{z}_i - E_O[\underline{z}_i|\underline{x}_i]\right)'\right] = E\left[\underline{z}_i\left(\underline{z}_i - E[\underline{z}_i|\underline{x}_i]\right)'\right] = V(\underline{z}_i|\underline{x}_i)$. In this case the bias term in the expression for b becomes

$$b = \beta - pV(\underline{z}_i)^{-1} V(\underline{z}_i|\underline{x}_i)\beta.$$

In the simplest case where z_i, x_i are scalars, this reduces to

$$b = \beta\left(1 - p\left(1 - R^2\right)\right).$$

Partition $\underline{z}_i$ into $\underline{z}_{1i}$ and $\underline{z}_{2i}$, where the function h is the identity function for the elements in $\underline{z}_{1i}$. This implies that $E_O[\underline{z}_{1i}|\underline{x}_i] = \underline{x}_{1i} = \underline{z}_{1i}$. Hence $(\underline{z}_{1i} - E_O[\underline{z}_{1i}|\underline{x}_i]) = 0$ for all i. Then, the matrix $\left(E_M\left[\underline{z}_i\left(\underline{z}_i - E_O[\underline{z}_i|\underline{x}_i]\right)'\right] - E[\underline{z}_i]\, E_M[\underline{z}_i - E_O[\underline{z}_i|\underline{x}_i]]'\right)$ can be written as

$$\begin{array}{cc} 0 & E_M\left[\underline{z}_{1i}\left(\underline{z}_{2i} - E_O[\underline{z}_{2i}|\underline{x}_i]\right)'\right] - E[\underline{z}_{1i}]\, E_M[\underline{z}_{2i} - E_O[\underline{z}_{2i}|\underline{x}_i]]' \\ 0 & E_M\left[\underline{z}_{2i}\left(\underline{z}_{2i} - E_O[\underline{z}_{2i}|\underline{x}_i]\right)'\right] - E[\underline{z}_{2i}]\, E_M[\underline{z}_{2i} - E_O[\underline{z}_{2i}|\underline{x}_i]]' \end{array}$$

The bias for b becomes

$$V(\underline{z}_i)^{-1}\left(\begin{array}{c} \left(E_M\left[\underline{z}_{1i}\left(\underline{z}_{2i} - E_O[\underline{z}_{2i}|\underline{x}_i]\right)'\right] - E[\underline{z}_{1i}]\, E_M[\underline{z}_{2i} - E_O[\underline{z}_{2i}|\underline{x}_i]]'\right)\beta_2 \\ \left(E_M\left[\underline{z}_{2i}\left(\underline{z}_{2i} - E_O[\underline{z}_{2i}|\underline{x}_i]\right)'\right] - E[\underline{z}_{2i}]\, E_M[\underline{z}_{2i} - E_O[\underline{z}_{2i}|\underline{x}_i]]'\right)\beta_2 \end{array}\right), \tag{15}$$

where β_2 is the corresponding partition of β. Hence the bias only depends on the prediction error in the regressor variables which differ in form from the corresponding match variables.

1.6 Imperfect Match on Multiple Categories

As noted in the text, we assume that $\underline{z}_i$ is a vector of $k-1$ binary variables representing k mutually exclusive categories (for example, educational categories). We assume that $x_i = 1$ represents the "last" J^* categories of $\underline{z}_i$ while $x_i = 0$ represents the reference category and the remaining categories of $\underline{z}_i$. Formally we define

$$x_i = \sum_{j \geq J^*} z_{ji}$$

where z_{ji} is the j^{th} element of $\underline{z}_i$. This structure implies that

$$\Pr[z_{ji} = 1 | x_i = 0] = \begin{cases} \frac{\Pr[z_{ij}=1]}{\Pr[x_i=0]} & \text{if } j < J^* \\ 0 & \text{otherwise} \end{cases}$$

and similarly

$$\Pr[z_{ji} = 1 | x_i = 1] = \begin{cases} \frac{\Pr[z_{ij}=1]}{\Pr[x_i=1]} & \text{if } j \geq J^* \\ 0 & \text{otherwise.} \end{cases}$$

It also follows that

$$E[y_i|x_i] = \alpha + \sum_{j=1}^{J^*-1} \Pr[z_{ji} = 1|x_i = 0]\,\beta_j + \left(\sum_{j=J^*}^{k-1} \Pr[z_{ji} = 1|x_i = 1]\,\beta_j\right) x_i.$$

Using the above expressions and the equation for the expectation of y_i,

$$\begin{aligned} E_s[y_i|x_i, \underline{z}_i] &= (1-p)\left(\alpha + \underline{z}_i'\underline{\beta}\right) \\ &\quad + p\left(\alpha + \sum_{j=1}^{J^*-1} \Pr[z_{ji} = 1|x_i = 0]\,\beta_j + \left(\sum_{j=J^*}^{k-1} \Pr[z_{ji} = 1|x_i = 1]\,\beta_j\right) x_i\right). \end{aligned}$$

Since $x_i = \sum_{j \geq J^*} z_{ji}$, this becomes

$$\begin{aligned} E_s\left[y_i|\underline{z}_i\right] &= \left(\alpha + p\sum_{j=1}^{J^*-1} \Pr\left[z_{ji}=1|x_i=0\right]\beta_j\right) \\ &+ \sum_{j=1}^{J^*-1} z_{ji}\left(1-p\right)\beta_j \\ &+ \sum_{j=J^*}^{k-1} z_{ji}\left(\left(1-p\right)\beta_j + p\sum_{l=J^*}^{k-1} \Pr\left[z_{li}=1|x_i=1\right]\beta_l\right). \end{aligned}$$

1.7 Implementation

The terms in equation 14 are all estimable in sample. Hence, an estimable expression for (α, β) is available from equation 14:

$$\begin{aligned} \alpha &= a - pE\left[\underline{z}_i\right]' \left(E\left[\underline{z}_i\underline{z}_i'\right] - E\left[\underline{z}_i\right]E\left[\underline{z}_i'\right]'\right)^{-1} \left(E_M\left[\underline{z}_i\left(\underline{z}_i - E_O\left[\underline{z}_i|\underline{x}_i\right]\right)'\right] - E\left[\underline{z}_i\right]E_M\left[\underline{z}_i' - E_O\left[\underline{z}_i|\underline{x}_i\right]'\right]\right)\beta \\ &\quad + pE_M\left[\underline{z}_i' - E_O\left[\underline{z}_i|\underline{x}_i\right]'\right]\beta \end{aligned}$$

$$\beta = \left(I - p\left(E\left[\underline{z}_i\underline{z}_i'\right] - E\left[\underline{z}_i\right]E\left[\underline{z}_i\right]'\right)^{-1}\left(E_M\left[\underline{z}_i\left(\underline{z}_i - E_O\left[\underline{z}_i|\underline{x}_i\right]\right)'\right] - E\left[\underline{z}_i\right]E_M\left[\underline{z}_i' - E_O\left[\underline{z}_i|\underline{x}_i\right]'\right]\right)\right)^{-1} b. \tag{16}$$

Step 1: Use OLS to estimate (a, b) on the full sample (including imputations). Retain the inverse of the variance of $\underline{z}_i$.

Step 2: Using the $R_i = O$ (observed) subsample, estimate $E_O\left[\underline{z}_i|\underline{x}_i\right]$. As a practical matter in CPS, this can be done using OLS on a full set of interaction terms for the imputation categories: age, education, gender, race, etc. Alternatively, this can be done by constructing all imputation cells and averaging within cell.

Step 3: Predict $\underline{z}_i$ using the estimated $E_O\left[\underline{z}_i|\underline{x}_i\right]$, for all observations in the $R_i = M$, sample (using the appropriate $\underline{x}_i$ for each observation).

Step 4: Construct $\underline{z}_i\left(\underline{z}_i - E_O\left[\underline{z}_i|\underline{x}_i\right]\right)'$ and $\left(\underline{z}_i - E_O\left[\underline{z}_i|\underline{x}_i\right]\right)$ in the $R_i = M$ sample, and average over that sample.

Step 5: p is estimated by the missing rate in the sample.

Step 6: Use estimated terms to construct estimates of α and β using equation 16.