Department of Economics
SOAS University of London

Working Paper Series

ISSN 1753 - 5816

Please cite this paper as:

Qin, Duo, Sophie van Huellen, Raghda Elshafie, Yimeng Liu, Thanos Moraitis. (2019), "A Principled Approach to Assessing Missing-Wage Induced Selection Bias", SOAS Department of Economics Working Paper Series, No. 216, London: SOAS University of London

No. 216

A Principled Approach to Assessing Missing-Wage Induced Selection Bias

by

Duo Qin, Sophie van Huellen, Raghda Elshafie, Yimeng Liu and Thanos Moraitis

(January 2019)

Department of Economics
SOAS University of London
WC1H 0XG
Phone: + 44 (0)20 7898 4730
Fax: 020 7898 4759
E-mail: economics@soas.ac.uk

A Principled Approach to Assessing Missing-Wage Induced Selection Bias

Duo Qin*
Department of Economics, SOAS University of London

Sophie van Huellen
Department of Economics, SOAS University of London

Raghda Elshafie
The Center for Victims of Torture

Yimeng Liu
School of Economics and Resource Management, Beijing Normal University

Thanos Moraitis
Department of Economics, SOAS University of London

Abstract

Multiple imputation (MI) techniques are applied to simulate missing wage rates of non-working wives under the missing-at-random (MAR) condition. The assumed selection effect of the labour force participation decision is framed as deviations of the imputed wage rates from MAR. By varying the deviations, we assess the severity of subsequent selection bias in standard human capital models through sensitivity analyses (SA). Our experiments show that the bias remains largely insignificant. While similar findings are possibly attainable through the Heckman procedure, SA under the MI approach provides a more structured and principled approach to assessing selection bias.

Keywords: wage, labour supply, selection, missing at random, multiple imputation

JEL classification: C21, C52, J20, J24

* SOAS University of London, Thornhaugh Street, Russell Square, London WC1H 0XG. Email: dq1@soas.ac.uk

1. Introduction

Missing data can pose serious challenges to inferential validity when inference is based on the available parts of the data samples. A prominent case in the context of empirical labour economics is missing wage rates for non-working wives when investigating the labour supply behaviour of married women. If married women's labour force participation (LPF) decision is regarded as a sample selection issue, any labour cost related inference drawn from subsample OLS estimates of working wives is deemed invalid as the OLS estimator suffers from selection bias (SB). A prominent solution to the SB problem is the two-stage Heckman procedure which augments the labour supply model by an inverse Mill's ratio obtained from a binary LFP selection equation; see Heckman (1974; 1976) and also Vella (1998) for the subsequent developments.

The Heckman procedure has become the standard approach, despite lack of conclusive empirical evidence of its correction for SB. Specifically, although the inverse Mill's ratio is found to be statistically significant more often than not, the consequent bias correction on the OLS coefficient estimates of explanatory variables of the labour supply models is frequently negligible; e.g. Moffitt (1999), Blau and Kahn (2007) and Van der Klaauw (2014). Furthermore, the significance of the inverse Mill's ratio is shown to be closely related to collinearity with selected covariates, raising questions regarding robustness of this ratio serving as evidence of the significant presence of SB; e.g. Moffitt (1999) and Puhani (2000). A study by Breuning and Mercante (2010) adds further doubt about the estimation accuracy of the Heckman procedure. When comparing predicted wage rates for non-employed individuals with observed wage rates when the same individuals re-enter employment, they find that OLS-based predictions consistently outperform predictions made by SB correction methods.

Seeking general explanations to the above accrued evidence, we reflect on the estimator-based route to correct SB from a methodological angle. The Heckman procedure was adapted from the tobit estimator which deals with truncated variables. We argue that it is implausible to regard missing wage rates of non-working wives as a data truncation problem. If a non-working wife has the same educational attainment and skills as a working wife, her shadow wage rate should be comparable to what the working wife is paid. If most of the subsample of non-working wives has earning attributes comparable to the subsample of working wives, what we learn from the latter group concerning their wage cost effect in labour supply models should be inferable to the former group. Indeed, if we consider the standard human capital model

explaining the wage rate, e.g. the Mincer model, all the explanatory variables are observed for both non-working and working wives and they capture conditions that exist before married women decide whether or not to join the labour force. This suggests that their earning potential might not be dependent on the LFP decision and it is non sequitur to assume SB based on that decision.

These considerations lead us to re-examining the missing wage rate issue by the multiple imputation (MI) approach pioneered by Rubin (1976); see also Little and Rubin (1987). Under this approach, SB amounts to asserting that the missing wage rates are *missing not at random* (MNAR). Following the MI approach, the assumed MNAR condition can be systematically investigated as deviation from stochastically imputed wage rates under the *missing at random* (MAR) condition via sensitivity analysis (SA). This way, the severity of SB resulting from MI wage rates under various plausible MNAR scenarios can be empirically assessed. The suggested approach is, to the best of our knowledge, unprecedented in the literature. Petreski *et al* (2014) come close in spirit. Their study uses MI to construct wage rates for non-working women to assess the gender wage gap for the Macedonian labour market. However, they do not utilise the MI approach to assess the severity of SB.

Our MI-based SA is carried out on a standard human capital model[1] using data from two widely used US-based data sources: The March Annual Demographic Survey of the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID). The findings can be summarised in three points. Firstly, a significant SB can only be induced under MNAR scenarios which deviate from the MAR condition substantially, so much so that plausibility of those scenarios is extremely low. Secondly, the few cases in which a significant SB can be produced put into question whether these should still be treated as SB in coefficients under a single population or as coefficients under different population classifications; see also Breuning and Mercante (2010) and Petreski *et al* (2014) on heterogeneity in the non-working subsample. Thirdly, the pattern of SB corrections by Heckman procedure is irregular, with most corrections confirming the finding of insignificant SB of the MNAR experiments. In short, our findings highlight the importance for labour economists to reorient their attention from concerns over SB to careful studies of the missingness mechanisms by means of various data matching tools when faced with missing wage rates.

[1] The application of SB correction estimators on the wage model has a long tradition, e.g. see Mroz (1987), and is still widely practised, e.g. Moeller (2002), Mercante and Mok (2014).

The paper is organised as follows. Following this brief introduction, section II introduces the MI approach and SA techniques. Section III describes experimental method and data used. Empirical results are reported and discussed in section IV and section V concludes with a brief discussion on methodological implications.

2. A Principled Assessment of SB via MI

The substantive model of our investigation into the severity of SB is a standard wage equation inspired by Mincer (1974):

$$(1) w_i = \boldsymbol{\beta} X_i + \varepsilon_i$$

with w_i being the logarithm of wage rate received by individual i, X_i being a covariate column vector comprising variables linked to human capital such as education, work experience and age, and possibly their polynomial forms, and $\boldsymbol{\beta}$ being a coefficient row vector.

When household survey data are used to estimate (1), the problem of incomplete samples arises with respect to the wage rate records. For example, when we examine (1) for married women, we find a subsample is not working and $\boldsymbol{\beta}$ as specified in (1) is inestimable. Let d_i be the missing data indicator with $d_i = 0$ if w_i is unobserved and $d_i = 1$ if w_i is observed. Further, let $w_{i,0} = w_i | d_i = 0$, $X_{i,0} = X_i | d_i = 0$ and $w_{i,1} = w_i | d_i = 1$, $X_{i,1} = X_i | d_i = 1$ and $w_{i,0+1} = w_{i,0} + w_{i,1}$. Given the missingness of $w_{i,0}$, we can only estimate the complete-case (CC) for the subsample of working wives:

$$(2) w_{i,1} = \boldsymbol{\beta}_1 X_{i,1} + \varepsilon_{i,1}, \quad \text{when } d_i = 1$$

It is widely accepted that $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}$ when the OLS estimator is used due to sample SB. The bias is embodied by an assumption of residual correlation, $corr(\varepsilon_i e_i) \neq 0$, between (1) and the following LFP decision model:

$$(3) d_i = \boldsymbol{\varphi} X_i + \boldsymbol{\theta} Z_i + e_i$$

with Z_i being another set of covariates pertinent to the LFP decision, e.g. husbands' earning and number of young children. However, $corr(\varepsilon_i e_i) \neq 0$ cannot be validated since (1) is inestimable.

To circumvent this impasse, we adopt the MI approach pioneered by Rubin (1976). At the core of this approach is the taxonomy of missing data mechanism, MAR when $Pr(d_i | w_0) =$

$Pr(d_i|w_i)$ versus MNAR when $Pr(d_i|w_0) \neq Pr(d_i|w_i)$. Notice that this probability-based condition underlies (3) in principle and that the SB assertion amounts to MNAR. This recognition leads us to utilise SA as part of the MI approach to investigate the severity of SB under different MNAR scenarios, e.g. see Carpenter and Kenward (2013, Ch10). Specifically, we produce various sets of stochastically simulated w_0 under MNAR as departures from stochastically simulated w_0 via MI under the MAR condition, such that the consequent SB in $\boldsymbol{\beta}_1$ under various plausible MNAR scenarios can be empirically assessed in a principled fashion. Hence, we first simulate w_0^{MAR} to generate a synthetic full-sample w_{0+1}^{MAR} by stacking w_0^{MAR} and w_1 together, and then produce different versions of w_{0+1}^{MNAR} from w_{0+1}^{MAR} as judged by the following alternative to (3):

(3') $\quad d_i = \alpha w_{0+1}^{MAR} + \boldsymbol{\varphi} X_i + \boldsymbol{\theta} Z_i + e_i, \qquad$ with $\alpha = 0$

$d_i = \alpha w_{0+1}^{MNAR} + \boldsymbol{\varphi}' X_i + \boldsymbol{\theta}' Z_i + e_i', \qquad$ with $\alpha \neq 0$

These various synthetic sets of w_{0+1}^{MNAR} can be used to empirically assess the severity of SB in $\boldsymbol{\beta}_1$ of (2).

It should be noted that model (2) precedes the missingness mechanism implied in (3'). Hence, there are no conclusive reasons that prohibit us from 'predicting' w_0 based on (2) by means of MI under the assumption of MAR. In view of the prevailing evidence of relatively low fits from empirical wage model studies using household survey data, we choose to impute w_0 by the predictive mean matching (PMM) method, see Little (1988). As a generalised hot deck method, PMM does not need any posterior distribution assumptions about ε_i; for the features of PMM and its empirical popularity see Carpenter and Kenward (2013: section 6.3), Morris *et al* (2014), Beretta and Santaniello (2016) and Murray (2018).

Under PMM, fitted and predicted values of w_1 and w_0 respectively are obtained from the CC estimation of a predictive wage rate model by the OLS. The predictive distance $D(i,j) = \widehat{w}_{i,0} - \widehat{w}_{j,1}$ with $i \neq j$ is then used to match each missing entry with its nearest neighbours such that the average wage rates of those neighbours are assigned as the imputed wage rates of those entries. The imputation is repeated m times, generating m version of $w_{i,0}^{MAR}$, following the Markov chain Monte Carlo (MCMC) principle. We can thus construct m sets of synthetic full-sample wage rates, $w_{i,0+1}^{MAR}$.

A vital requirement of predictive models for the MI purpose is to include variables predictive of the missingness mechanism; see Murray (2018), Carpenter and Kenward (2013: Ch. 2). Such auxiliary variables have already been introduced as Z_i in (3). Hence, we use the following augmented wage model (4) as our basic predictive model for MI:

$$(4) w_{i,1} = \boldsymbol{b}_1 X_{i,1} + \boldsymbol{\gamma}_1 Z_{i,1} + v_{i,1}, \qquad \text{when } d_i = 1$$

It should be noted that model (4) might be a more accurate wage model for married women than the standard human capital model, as the wage rate acceptable to working wives may depend on Z_i. The model is closer to the shadow wage model by Heckman (1974).

However, model (4) might still be miss-specified, or the OLS estimates inconsistent, for the MI purpose. Two 'doubly robust' methods are thus adopted here: (a) inverse probability weighting (IPW) and (b) a doubly robust nonparametric (DRN) method; see Murray (2018). Both methods augment and modify the role of model (4) in MI by the predicted probability from model (3).

IPW modifies the OLS applied to (4) by weighting each observation with the inverse of probability π_i of it having been observed. Individuals that are less likely to be observed and hence more similar to those missing, are thereby given a larger weight. The *p*-scores from model (3) are used as estimates for π_i. As a result, model (4) is modified into:

$$(4') \qquad w_{i,1} = \boldsymbol{b}_{1\pi} X_{i,1} \left(\hat{\pi}_{i,1}\right)^{-1} + \boldsymbol{\gamma}_{1\pi} Z_{i,1} \left(\hat{\pi}_{i,1}\right)^{-1} + v_{i,1\pi}, \text{ when } d_i = 1$$

IPW/MI using (4') generates m sets of $w_{i,0+1\pi}^{MAR}$. The advantage of the IPW/MI method is that it is robust against miss-specification in either the predictive model or the logistic model, but not both; see Vansteelandt *et al* (2010), Carpenter *et al* (2006), Seaman and White (2013) and Seaman *et al* (2012).

The DRN/MI method utilises the *p*-scores from (3) in a non-parametric way, different from the IPW/MI method, see Long *et al* (2012) and Hsu *et al* (2016). The *p*-scores are used as part of the nearest neighbour selection process. Specifically, DRN/MI uses the fitted and predicted values from (4) and (3) to construct a set of composite scores $S = (\widehat{W}, \widehat{\Pi})$, where all the predicted values are standardised, indicated by use of capital letters. The distance $D_\omega(i,j)$ is used to find for each subject i with missing w_i the k nearest observed neighbours:

$$(5) D_\omega(i,j) = \left\{\omega_1\left[\widehat{W}_{i,0} - \widehat{W}_{j,1}\right]^2 + \omega_2\left[\widehat{\Pi}_{i,0} - \widehat{\Pi}_{j,1}\right]^2\right\}^{1/2}$$

with $\omega_1 + \omega_2 = 1$. Note that with $\omega_1 = 1$, DRN/MI is equivalent to the OLS/MI method. We follow Hsu *et al* (2016) in using a kernel-based version of the DRN/MI to construct m sets of $w_{i,0+1\omega}^{MAR}$.[2]

Having imputed three sets of w_0, i.e. by OLS/MI, IPW/MI and DRN/MI respectively under the MAR condition, we are in the position to design different scenarios of MNAR for SA based on (3') described earlier. With regard to the PMM-based MI approach adopted presently, we construct MNAR scenarios through simply imposing mean shifts on the imputed sets of w_0 under MAR by a sensitivity parameter, $\delta \neq 1$: $w_0^{MNAR} = \delta w_0^{MAR}$, following Rubin's (1987) original suggestion. The resulting w_0^{MNAR} sets are then stacked onto w_1 to form w_{0+1}^{MNAR}. These synthetic wage rate sets enable us to run model (1). Since our MI experiments via the three methods have generated a reasonably large set of imputed wage rates, we can pool them together to produce a pair of single $\overline{\boldsymbol{\beta}}_{0+1}^{MNAR}$ and $\overline{\boldsymbol{\beta}}_0^{MNAR}$ estimates by means of the Rubin's combination rule using the following variations of (1):

$$(6a) \qquad w_{i,0+1}^{MNAR} = \overline{\boldsymbol{\beta}}_{0+1}^{MNAR} X_{i,0+1} + \varepsilon_{i,0+1}^*,$$

$$(6b) \qquad w_{i,0}^{MNAR} = \overline{\boldsymbol{\beta}}_0^{MNAR} X_{i,0} + \varepsilon_{i,0}^*, \text{ when } d_i = 0.$$

The severity of SB can then be assessed by comparing these $\boldsymbol{\beta}$ estimates with those from model (2). This type of SA follows what is known as the 'pattern mixture' route, e.g. see Carpenter and Kenward (2013: Chapter 10). Our first MNAR scenario is aimed at identifying a 'just MNAR' case by (3'). Specifically, we conduct a tipping point analysis to search for δ by gradually altering it in the neighbourhood of 1, such that the resulting δw_0^{MAR} will lead to $\alpha \neq 0$ in (3'); see Liublinska and Rubin (2012). Our other MNAR scenarios are aimed at assessing how the severity of SB could be aggravated through δ values further away from 1. To set these values within plausible boundaries, we look into the wage data features and decide on two scenarios: one utilising the minimum wage rates and the other the average wage rates of husbands. The detailed design is described in the next section.

[2] We are grateful to Hsu *et al* (2016) for provision of their source code which we adapted for this paper.

3. Data and Method

The empirical analysis is carried out for five years, 1981, 1991, 2001, 2011 and 2015 using samples from PSID and CPS data sources. Years of education and age are two common variables constituting X_i in equation (1). While age is comparable across the two data sources, the coding for the education variable differs, which hampers comparability. For the PSID samples, years of previous work experience is available and included in X_i. Unfortunately, this variable is unavailable from the CPS source. Age is therefore used as a proxy of work experience. Z_i in (3) is formed of three variables available via both data sources: husband's wage rate, number of children, and a binary variable indicating if any of the children's age is under six. A detailed description of the two datasets and processing of the data can be found in Appendix 1.

Roughly a quarter of wives is not in the labour force. While the proportion of non-working wives is fairly constant across years and data sources, the average wage rates received by working women increase notably over the years. In 1981, average wage rates are around $6, almost double the minimum wage of $3.1, and rise to $24 in 2015, while the minimum wage only reaches $7.25. This rise is accompanied by a rapid narrowing of the gap between wife and husband mean wage rates. In 1981, the average husbands' wage rate is around 1.7 times higher than the average wage rate received by working women. The ratio drops to roughly 1.5 in 2001 and shrinks further to 1.3 in the last two years. These changes are accompanied by a notable rise in the educational attainment of working wives.

A close comparison of working wives with non-working wives reveals that the key differences between the two sub-samples lie in educational attainment, the number of children in the household and the presence of young children in the household. These three variables serve as important predictors of the missingness in the wage rate variable and hence are essential regressors in the selection model (4).

In the empirical specification of (4), we follow the common practice of using polynomial and cross-product terms of the variables in X_i to capture possible nonlinear effects in the wage model. Specifically, the initial X_i contains not only education and age (and previous work experience in the case of PSID sets), but also their quadratic and cubic terms as well as their cross products. Correlation analysis of all possible regressors shows that correlations between X_i and Z_i are low, indicating a low risk of omitted variable bias between the beta estimates of

(2) and (4). In contrast, correlations of variables in X_i and their polynomial terms are very high; all above 0.9 and some even above 0.95. The estimated effects of education or age should hence not be interpreted singularly but in combination with their polynomial terms when present. Backward model selection is used to select a parsimonious model from a general model specification of (4), which results in slightly different model specifications across different years. To facilitate cross data source model comparison, we try to minimise specification differences between the two data sources within the same year, provided that this endeavour does not go against the backward model selection principle. The same criteria are applied to the selection of the logistic model (3).

Since the null hypothesis $\alpha \neq 0$ via (3') in SA is false by the MNAR design, it is inadequate to judge $\alpha \neq 0$ solely by the commonly used 5 per cent significance level without looking into the power of the test, or the probability of type II error. Trial calculation of this probability corresponding to the 5 per cent significance level shows a virtually zero value, thanks to our relatively large sample sizes. Hence, the null-hypothesis based standard criterion for significance still works here.

However, the unequal sample size complicates cross data source comparability, when it comes to search for the tipping point, denoted by δ_{tp}, in the SA. The CPS data sets are about ten times larger than their PSID counterparts. The different sample size may undermine consistent choice of δ_{tp} by p-values. To accommodate for this difference in sample size, we rely on effect size estimates as a measure independent of sample size (Lin *et al* 2013; Kelley and Preacher 2012). In particular, we run the tipping point analysis with PSID data sets first to find the effect size estimate (odds ratio) $exp(\hat{\alpha}_S)$ by logistic regression estimation of (3') for which $\delta_{tp,S}$ yields 'just MNAR'. We then choose $\delta_{tp,L}$ for CPS data sets so that $exp\ (\hat{\alpha}_S) \approx exp\ (\hat{\alpha}_L)$ as a comparable sensitivity parameter to $\delta_{tp,S}$ of the smaller sample.

The tipping point analysis only allows us to examine the sensitivity of SB at a marginal situation where MAR is just being violated. In order to consider economically plausible situations where the violations go beyond this 'just MNAR' scenario, we design two further scenarios: 'minimum wage MNAR' and 'husband wage MNAR'. The two scenarios are expected to capture opposite sides in terms of direction and strength of SB over the spectrum of feasible MNAR situations. Under the former scenario, the Federal minimum wage rate of $7.25, which was set in 2009 and remained effective for the 2011 and 2015 samples, is used as

the threshold value to construct and design δ_{mw}. Compared with the average wage rates of these two years, this threshold implies a wage reduction of 0.33% and 0.30% respectively.

It should be noted that these ratios are not directly convertible into a single δ_{mw} that could be multiplied with w_0^{MAR} due to the logarithm model form. We thus derive the SA parameter via applying these wage reduction ratios to the mean sets of the MI wage variable.[3] For the years before 2009, we reduce δ_{mw} step wise towards δ_{tp} (see Table 7). We are thereby able to examine a range of sensitivity results as the severity of MNAR situations increases towards the minimum wage threshold. Under the 'husband wage' scenario, the average of husband wage rate for each cohort is used to derive δ_{hw}, in the same way as \$7.25 is used to derive δ_{mw} for 2011 and 2015 (see footnote 3). Furthermore, we exploit the variations in the ratios of the average MI wage variable to introduce a stochastic element to the mean-shift based simulations. Specifically, δ_{hw} becomes randomised:

$$(7)\ r_{i,0} = \frac{\bar{w}_{i,0}^{MAR} + ln(\overline{W_h})}{\bar{w}_{i,0}^{MAR}}, \qquad \delta_{hw} : \sim N\left(\overline{r_{i,0}}, S_{r_{i,0}}\right)$$

where $\overline{W_h}$ denotes husband's average wage rate, $\overline{r_{i,0}}$ and $S_{r_{i,0}}$ are the mean and standard deviation of $r_{i,0}$ respectively (see Tables A2 and 7 for the actual values used).

To evaluate the severity of SB, the $\boldsymbol{\beta}_{0+1}^{MNAR}$ and $\boldsymbol{\beta}_{0}^{MNAR}$ estimates obtained via (6) under the three different SA scenarios are compared against $\boldsymbol{\beta}_1$ from (2) and their 95% confidence intervals.[4] For consistency, the model specification of (2) matches (4) in the choice of covariates X_i. In addition, we also provide an estimate, denoted as $\boldsymbol{\beta}_{1H}$, by the standard Heckman 2-step procedure. Specifically, this is obtained via augmenting (2) with a covariate, known as the inverse Mill's ratio, which is derived from (3) under the assumption that $\rho = corr\left(\varepsilon_{i,1} e_i\right) \neq 0$. We start from the chosen (3) for the MI experiments as described above and revise them to make sure that the exclusion restrictions are satisfied as required by the Heckman procedure.

[3] Take 2011 as an example, $r_{i,0} = \frac{\bar{w}_{i,0}^{MAR} + ln(0.33)}{\bar{w}_{i,0}^{MAR}}$, $\delta_{mw} := \overline{r_{i,0}} = 0.6$, for the construction of $\bar{w}_{i,0+1\delta}^{MNAR}$.

[4] We are aware that the MI variance estimator by Rubin's combination rule becomes inconsistent under improper imputations; see Carpenter and Kenward (2013: pp.62), and Murray (2018). Xie and Meng (2017) suggest using 2*MI Rubin's variance as a simple adjustment instead. This adjustment would result in a less severe SB in our case. Nevertheless, Reiter (2017) notes that this inconsistency is of much less practical importance than coefficient bias. We hence refrain from the adjustment.

To summarise, the above steps provide us with up to seven $\boldsymbol{\beta}$ estimates under different MNAR scenarios per year and data source to compare the CC estimate $\boldsymbol{\beta}_1$ with $\overline{\boldsymbol{\beta}}^{MNAR}_{0+1,\delta_{tp}}, \overline{\boldsymbol{\beta}}^{MNAR}_{0,\delta_{tp}}$, $\overline{\boldsymbol{\beta}}^{MNAR}_{0+1,\delta_{mw}}, \overline{\boldsymbol{\beta}}^{MNAR}_{0,\delta_{mw}}, \overline{\boldsymbol{\beta}}^{MNAR}_{0+1,\delta_{hw}}, \overline{\boldsymbol{\beta}}^{MNAR}_{0,\delta_{hw}}$ and $\boldsymbol{\beta}_{1H}$ respectively.

4. Empirical Results

Table 1 reports the coefficient estimates of the two X_i variables of interest from a SB point of view, education and age, and their quadratic terms from the CC OLS regression estimation of model (4). Although not reported here, most of the coefficient estimates of the covariates in Z_i are found highly significant, justifying their inclusion in the model. Further, model specification search for (4) results in the retained polynomial and cross-product terms in X_i across years, a clear case of non-linear forms. The relatively low fit as reflected by the low R^2 reported and the overwhelming rejection of the normality assumption of the residuals by the Shapiro-Francia normality test lend further support to our choice of PMM as the stochastic residual simulation method in our MI experiment.

Table 1. $\boldsymbol{\beta_1}$ Estimates for Education and Age Variables and R-squares from the CC Estimation of (4)

	Education		Education²		Age		Age²		S-F Normality		R^2
	PSID	CPS	PSID	CPS	PSID	CPS	PSID	CPS	PSID	CPS	PSID /CPS
1981	0.0769** (.0069)	-0.1783** (.02758)	n/a	0.0029** (.00035)	0.0234 (.0148)	0.0157** (.00440)	-0.0003 (.00018)	-0.0002** (.00005)	0.9644 (0.000)	0.9685 (0.000)	0.2166 /0.11
1991	-0.0646** (.01696)	-0.3167** (.03254)	0.0063** (.00074)	0.0049** (.00041)	-0.0063** (.00185)	0.0376** (.00466)	n/a	-0.0004** (.00006)	0.9732 (0.000)	0.9739 (0.000)	0.2682 /0.2084
2001	0.0851** (0.0056)	-0.3450** (.03316)	n/a	0.0054** (.00042)	0.0597** (.0121)	0.0377** (.00534)	-0.0007** (0.0002)	-0.0004** (.00006)	0.9829 (0.000)	0.9752 (0.000)	0.2734 /0.2196
2011	0.0980** (.00582)	-0.2919** (.03118)	n/a	0.0050** (.00037)	0.0374** (.01243)	0.0522** (.00847)	-0.0003* (.00015)	-0.0004** (.00005)	0.9803 (0.000)	0.9777 (0.000)	0.2123 /0.2343
2015	-0.1067* (.05433)	-0.3080** (.03686)	0.0063** (.00141)	0.0054** (.00044)	0.0333* (.01514)	0.0626** (.00974)	-0.0005** (.00014)	-0.0003** (.00006)	0.9834 (0.000)	0.9833 (0.000)	0.2504 /0.2343

Notes: The coding of the education variable differs between PSID and CPS samples and coefficients estimates are therefore not expected to align. ² indicates the quadratic term of the education and age variable. Standard errors in (.). * indicates significance at the 5% and ** at the 1% level respectively. 'S-F Normality' is the test statistic of the Shapiro-Francia normality test and respective p-value.

Prediction accuracy of the *p*-score from logistic estimation of model (3) is assessed by means of the receiver operating characteristic (ROC) in Table 2. The *p-score* prediction of working women is very high, with a sensitivity statistic of around 90 per cent and higher. These statistics contribute to the significant ROC area and lend support for the use of *p-score* estimates of the working women group in the IPW/MI. The weights used for the DRN/MI in (5) are set to $\omega_1 = 0.6$ for model (4) and $\omega_2 = 0.4$ for model (3) after extensive experiments

on the weight variations, judged mainly by the relative variance increase (RVI) and the fraction of missing information (FMI) statistics of MI.

Table 2. Prediction Accuracy of (3) and Estimated ROC Area

	Sensitivity (%)		Specificity (%)		ROC area (s.d.)	
	PSID	CPS	PSID	CPS	PSID	CPS
1981	95.98	89.95	17.89	25.88	0.7178 (0.0115)	0.6880 (0.0037)
1991	97.20	96.78	21.81	11.72	0.7661 (0.0101)	0.6939 (0.0042)
2001	99.57	97.88	2.31	8.63	0.6962 (0.0140)	0.6813 (0.0049)
2011	99.38	97.55	4.24	10.62	0.6879 (0.0133)	0.6660 (0.0040)
2015	99.78	96.60	3.38	14.09	0.6967 (0.0133)	0.6775 (0.0041)

Note: The classification statistics are based on the threshold value of 0.50 for the predicted p-scores.

After various trial experiments on the numbers of imputation m and neighbours k in PMM, we set $k = 4$ for the smaller PSID samples and $k = 10$ for the large CPS samples.[5] The number of imputations is set at $m = 50$ in consideration of the FMI statistics reported in Table 3.

Table 3. RVI and FMI Statistics for Education and Age Variables from PMM OLS/MI

		Education		Education^2		Age		Age^2		Av RVI/max FMI	
		PSID	CPS	PSID	CPS	PSID	CPS	PSID	CPS	PSID	CPS
		Relative Variance Increase (RVI)									
1981	OLS	0.3135	0.5167	n/a	0.4900	0.3816	0.5094	0.4226	0.5484	0.3896	0.5267
	IPW	0.2708	0.4767	n/a	0.4591	0.4152	0.9587	0.449	1.0543	0.3709	0.596
	DRN	0.6492	0.9231	n/a	0.8672	0.5524	0.6867	0.6093	0.7795	0.5314	0.7267
1991	OLS	0.6577	0.4040	0.5342	0.3868	0.4535	0.2919	n/a	0.2890	0.3309	0.3929
	IPW	0.7241	0.3929	0.6089	0.3864	0.4735	0.4895	n/a	0.5011	0.3322	0.4204
	DRN	0.9296	0.9723	0.7103	0.9205	0.7846	0.4579	n/a	0.4989	0.5388	0.5604
2001	OLS	0.2368	0.3171	n/a	0.3063	0.2170	0.2645	0.2404	0.2664	0.1973	0.3571
	IPW	0.2151	0.3601	n/a	0.3496	0.2451	0.2867	0.2697	0.3014	0.2133	0.3497
	DRN	0.2354	0.5977	n/a	0.5615	0.2739	0.2614	0.3151	0.2805	0.2603	0.5088
2011	OLS	0.3535	0.4032	n/a	0.4544	0.2332	0.4374	0.2491	0.3029	0.2513	0.3918
	IPW	0.2883	0.3349	n/a	0.3704	0.1957	0.2977	0.2095	0.4244	0.2555	0.3694
	DRN	0.1834	0.6779	n/a	0.5574	0.1942	0.6884	0.2104	0.3233	0.2804	0.5704
2015	OLS	0.5488	0.6415	n/a	0.6214	0.2452	0.4033	0.2603	0.5690	0.3002	0.4733
	IPW	0.3942	0.4699	n/a	0.5469	0.2248	0.7151	0.1383	0.3793	0.2748	0.4595
	DRN	0.7153	1.2063	n/a	0.9972	0.3669	1.0357	0.3130	0.6172	0.3283	0.6747
		Fraction of Missing Information (FMI)									
1981	OLS	0.2406	0.3438	n/a	0.3318	0.2786	0.3406	0.2998	0.3575	0.3435	0.4133
	IPW	0.2147	0.3257	n/a	0.3174	0.2961	0.4944	0.3128	0.5184	0.3573	0.5184
	DRN	0.3978	0.4849	n/a	0.4691	0.3594	0.4111	0.3825	0.4424	0.4159	0.4874
1991	OLS	0.4008	0.2902	0.3516	0.2812	0.3149	0.2276	n/a	0.2258	0.4008	0.4000
	IPW	0.4243	0.2844	0.3823	0.2810	0.3244	0.3316	n/a	0.3369	0.4243	0.3369
	DRN	0.4869	0.4980	0.4196	0.4842	0.4443	0.3168	n/a	0.3359	0.4869	0.5099
2 0	OLS	0.1928	0.2426	n/a	0.2362	0.1795	0.2106	0.1952	0.2118	0.1952	0.3455

[5] Experiments with different k did not make much of a difference in the case of CPS data sets. However, for PSID, larger k resulted in a slightly narrower distribution of imputed wage rates, but the overall impact of different k values on the distribution is quite small as long as $k = 1$ is disregarded, e.g. Beretta and Santaniello (2016).

	IPW	0.1782	0.2669	n/a	0.2611	0.1983	0.2244	0.2140	0.2333	0.2140	0.3273
	DRN	0.1919	0.3777	n/a	0.3630	0.2166	0.2086	0.2415	0.2206	0.2504	0.3862
2011	OLS	0.2634	0.2897	n/a	0.3152	0.1904	0.3069	0.2009	0.2342	0.2634	0.3693
	IPW	0.2255	0.2528	n/a	0.2725	0.1647	0.2311	0.1743	0.3005	0.2255	0.3445
	DRN	0.1559	0.4080	n/a	0.3613	0.1636	0.4117	0.1750	0.2462	0.2783	0.4544
2015	OLS	0.3579	0.3946	n/a	0.3869	0.1983	0.2898	0.2080	0.3661	0.3749	0.3946
	IPW	0.2853	0.3225	n/a	0.3568	0.1848	0.4211	0.1221	0.2773	0.2975	0.4211
	DRN	0.4214	0.5522	n/a	0.5044	0.2707	0.5139	0.2403	0.3853	0.4392	0.5783

Notes: The coding of the education variable differs between PSID and CPS samples and coefficients estimates are therefore not expected to align. ² indicates the quadratic term of the education and age variable. 'Av RVI' is the average RVI over all covariates. 'max FMI' is the maximum FMI over all covariates.

Computing time was an additional factor considered, especially with respect to the DRN/MI method. It is noticeable that the RVI statistics due to missingness tend to rise as we move from OLS/MI to IPW/MI and DRN/MI. This reflects the expected cost for consistency by sacrificing efficiency. After each set of MIs, the logistic regression (3') is run to verify $\alpha \neq 0$ and hence MAR.

In order to check whether models (4), (3) and (4') provide adequate neighbours for PMM imputation, we examine the minimum and the maximum of the fitted values of these models from the working women subsample and compare them with the predicted values of the non-working wife subsamples in Table 4. When the latter falls outside the range of the former, the number of outliers in the non-working wife subsamples are calculated. We see that such cases are rare, and outliers are very few if they exist, indicating that most of the non-working wives find potential matches from the working wife group.[6]

Table 4. Ranges of $\widehat{w}_0$, $\widehat{w}_1$, and P-scores From Estimation of (4), (3) and (4')

		Fitted from OLS of (4)				P-score from Logit of (3)				Fitted from WLS of (4')			
		PSID		CPS		PSID		CPS		PSID		CPS	
		Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
1981	$d=0$	0.2767	2.5701	0.8044	2.2073	0.1443	0.9446	0.0814	0.9571	0.2487	2.5732	0.7755	2.2008
	$d=1$	0.2636	2.5715	0.8720	2.2476	0.1205	0.9687	0.0815	0.9744	0.2548	2.5612	0.8328	2.2415
	Outlier	0	0	5	0	0	0	1	0	1	1	4	0
1991	$d=0$	0.9558	3.1423	1.1943	2.9410	0.0308	0.9743	0.1032	0.9701	0.9679	3.1537	1.1662	2.9417
	$d=1$	1.0140	3.1005	1.2167	3.1318	0.1120	0.9866	0.0966	0.9803	1.0145	3.1105	1.2074	3.1259
	Outlier	2	3	2	0	12	0	0	0	2	3	2	0
2001	$d=0$	0.8316	3.3696	1.5752	3.4892	0.2124	0.9894	0.1106	0.9540	0.8015	3.3807	1.5700	3.4894
	$d=1$	0.7674	3.5911	1.5332	3.4879	0.2057	0.9905	0.1818	0.9721	0.7423	3.5978	1.5276	3.4875
	Outlier	0	0	0	1	0	0	3	0	0	0	0	1
2011	$d=0$	1.3918	3.5360	1.6328	3.8581	0.1108	0.9635	0.0933	0.9514	1.3499	3.5449	1.6135	3.8582
	$d=1$	1.4674	3.6715	1.8690	3.7883	0.2730	0.9768	0.1086	0.9644	1.4438	3.6847	1.8498	3.7859
	Outlier	3	0	3	2	5	0	2	0	3	0	3	2
2015	$d=0$	1.9801	3.8276	1.8759	3.8926	0.2840	0.9707	0.0656	0.9502	1.9970	3.8315	1.8700	3.8980
	d=1	1.9457	3.8554	1.8176	3.8499	0.3647	0.9868	0.0763	0.9560	1.9513	3.8609	1.8140	3.8519
	Outlier	0	0	0	1	3	0	1	0	0	0	0	1

Notes: 'Min' is the minimum and 'Max' the maximum wage rate/p-score from the fitted observed ($d=1$) and predicted missing ($d=1$) set obtained from fitting (4) and (5). 'WLS' is the weighted least square using IPW on (4). 'Outlier' is the number of predicted values falling outside the min/max range of the fitted values.

[6] We ran similar checks using the standard wage model (2) for MI instead of (4). Unsurprisingly, the number of outliers is distinctly larger than reported in Table 4. This shows the importance of including covariates Z_i that are predictive of missingness in the predictive model, as already discussed in the methodological section.

Table 5 provides basic summary statistics of the MI wage rates compared to those of the observed wage rates. Distributions of MI wage rates are more centred with distinctly smaller standard deviations than those of the observed wage rates. In contrast, there is far less dissimilarity between distributions of pairwise comparisons of the three sets of imputed wage rates. Kolmogorov-Smirnov test of equality of distributions reported in Table 6 show that there is no statistical difference between the three sets of MIs from the PSID data sets, and little difference for the CPS sets.[7]

Table 5. Distribution of w_1 and w_0^{MAR} Under Different MI Methods

		PSID				CPS			
		w_1	w_0^{MAR}			w_1	w_0^{MAR}		
		Observed	OLS/MI	IPW/MI	DRN/MI	Observed	OLS/MI	IPW/MI	DRN/MI
1981	Mean	1.6659	1.5628	1.5637	1.5578	1.5764	1.5350	1.5328	1.5368
	s.d.	0.6243	0.3024	0.3053	0.3311	0.5473	0.1995	0.2016	0.2222
	Min	-0.6162	0.1396	0.1255	0.3161	-0.6035	0.8192	0.7841	0.6520
	Max	4.6051	2.5120	2.4396	2.3836	4.9053	2.4242	2.4241	2.7163
1991	Mean	2.1410	1.9111	1.9126	1.9141	2.1337	2.0498	2.0468	2.0519
	s.d.	0.6747	0.4227	0.4228	0.4226	0.6064	0.2916	0.2941	0.3050
	Min	-0.1508	0.8630	0.8819	0.7360	-0.1823	1.2618	1.2715	0.9489
	Max	4.6051	3.2046	3.2995	3.3418	5.3391	3.1168	3.1611	3.0982
2001	Mean	2.5621	2.5048	2.5117	2.5126	2.5419	2.4626	2.4587	2.4617
	s.d.	0.6340	0.3878	0.3927	0.3990	0.6321	0.3356	0.3364	0.3490
	Min	0.0677	1.4569	1.5861	1.5905	0.0561	1.5523	1.4964	1.5224
	Max	5.0360	3.3605	3.4067	3.3739	5.7446	3.6922	3.7990	3.7217
2011	Mean	2.8569	2.7633	2.7638	2.7569	2.8743	2.7774	2.7784	2.7699
	s.d.	0.6719	0.3633	0.3635	0.3627	0.6516	0.3645	0.3641	0.3750
	Min	0.3075	1.6582	1.7544	1.9025	0.3175	1.8691	1.8149	1.8183
	Max	5.4525	3.7720	3.6956	3.8574	5.9757	3.8481	4.0013	3.9868
2015	Mean	2.9462	2.8113	2.8128	2.7924	2.9641	2.8214	2.8240	2.8218
	s.d.	0.6554	0.3586	0.3571	0.3639	0.6693	0.3618	0.3652	0.3752
	Min	0.5008	2.0491	1.9060	1.9214	0.4055	1.8332	1.9362	1.7899
	Max	5.2478	3.7460	3.7038	3.7511	6.0715	4.0497	4.1211	4.3013

Note: MI statistics are based on the averages of $m = 50$.

Table 6. Equality of Distribution Test for w_0^{MAR} Under Different MI Methods

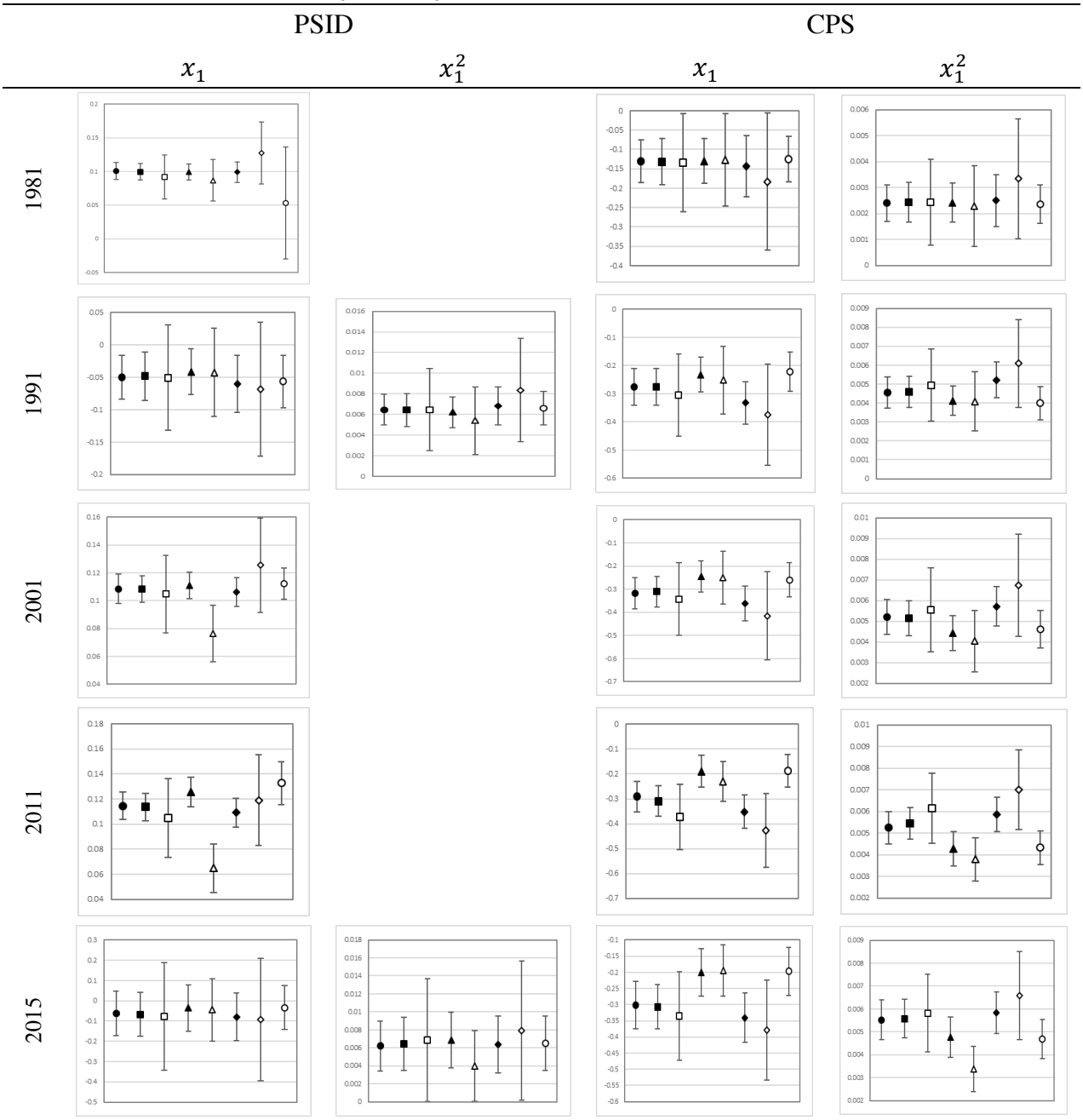
	OLS versus IPW		OLS versus DR		IPW versus DR	
	PSID	CPS	PSID	CPS	PSID	CPS
1981	0.0415	0.0154	0.0543	0.0353**	0.0591	0.0293**
	(0.653)	(0.318)	(0.314)	(0.000)	(0.224)	(0.003)
1991	0.0223	0.0133	0.0223	0.0236	0.0297	0.0319*
	(0.996)	(0.772)	(0.996)	(0.126)	(0.928)	(0.013)
2001	0.0282	0.0134	0.0436	0.0273	0.0410	0.0275
	(0.998)	(0.885)	(0.853)	(0.117)	(0.898)	(0.111)
2011	0.0254	0.0081	0.0466	0.0180	0.0403	0.0209
	(0.998)	(0.987)	(0.684)	(0.271)	(0.839)	(0.135)
2015	0.0383	0.0136	0.0450	0.0200	0.0541	0.0140
	(0.901)	(0.639)	(0.759)	(0.184)	(0.535)	(0.608)

[7] Kolmogorov-Smirnov test is run between observed wage rates and imputed wage rates and the results are all rejections. We decide not to report those test statistics but those summary statistics in Table 4 as those statistics are more telling than the K-S test statistics, which reject overwhelmingly that the imputed wages distribute similarly as those of observed wages.

Notes: Two-group Kolmogorov-Smirnov test of equality of distributions. MI statistics are based on the averages of $m = 50$. P-values in (.).

Figure 1a: $\boldsymbol{\beta}$ estimates from (6) with 95% confidence intervals for $x_1 = Education$

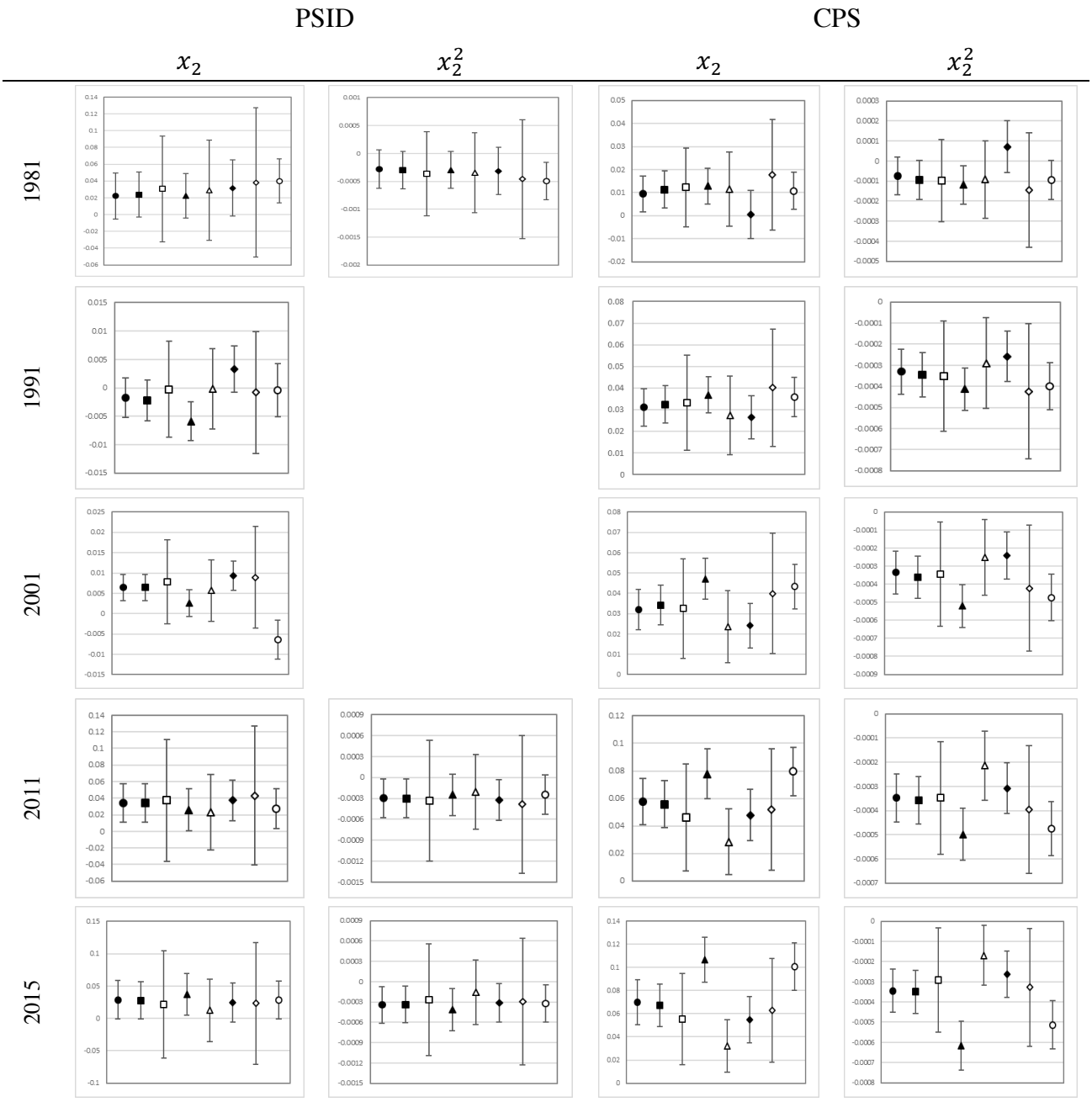
β_1: ● , β_{1H}: ○; $\bar{\beta}^{MNAR}_{0+1,\delta_{tp}}$: ■, $\bar{\beta}^{MNAR}_{0,\delta_{tp}}$: □; $\bar{\beta}^{MNAR}_{0+1,\delta_{mw}}$: ▲, $\bar{\beta}^{MNAR}_{0,\delta_{mw}}$: △; $\bar{\beta}^{MNAR}_{0+1,\delta_{hw}}$: ◆, $\bar{\beta}^{MNAR}_{0,\delta_{hw}}$: ◇



Note: The coding of the education variable differs between PSID and CPS samples and coefficients estimates are therefore not expected to align. ^{2} indicates the quadratic term of the education.

Figure 1b: $\boldsymbol{\beta}$ Estimates from (6) with 95% confidence intervals for $x_2 = Age$

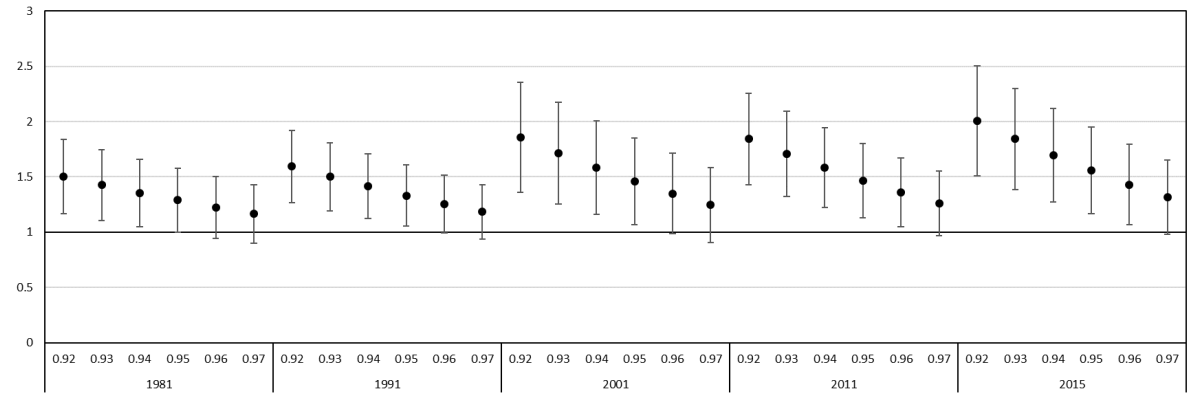
β_1: ● , β_{1H}: ○; $\bar{\beta}^{MNAR}_{0+1,\delta_{tp}}$: ■, $\bar{\beta}^{MNAR}_{0,\delta_{tp}}$: □; $\bar{\beta}^{MNAR}_{0+1,\delta_{mw}}$: ▲, $\bar{\beta}^{MNAR}_{0,\delta_{mw}}$: △; $\bar{\beta}^{MNAR}_{0+1,\delta_{hw}}$: ◆, $\bar{\beta}^{MNAR}_{0,\delta_{hw}}$: ◇



Note: ^{2} indicates the quadratic term of the age variable.

Core findings of the SA-based SB bias assessment are summarised in Figures 1a and 1b, which plot the estimated $\boldsymbol{\beta}_1$ of the two key variables – education, age, and their quadratic terms – and the corresponding $\overline{\boldsymbol{\beta}}^{MNAR}_{0+1,\delta_{tp}}, \overline{\boldsymbol{\beta}}^{MNAR}_{0,\delta_{tp}}$, $\overline{\boldsymbol{\beta}}^{MNAR}_{0+1,\delta_{mw}}, \overline{\boldsymbol{\beta}}^{MNAR}_{0,\delta_{mw}}, \overline{\boldsymbol{\beta}}^{MNAR}_{0+1,\delta_{hw}}, \overline{\boldsymbol{\beta}}^{MNAR}_{0,\delta_{hw}}$ under different MNAR scenarios, and also $\boldsymbol{\beta}_{1H}$, together with their 95% confidence intervals. Figure 2 further summarises the tipping-point search for δ_{tp} in the 'just MNAR' scenario via (3'). We find values for $\delta_{tp,S}$ around 0.95-0.97 across different years. As evident from Figure 2, $\delta_{tp,L}$ values for the CPS data sets found by matching the odd ratios of the estimated α in (3'), do not differ much from $\delta_{tp,S}$. This finding eases our concerns over cross data set comparability as the found variations in δ_{tp} are too small to warrant much practical concern. Further, the consequent bias in $\boldsymbol{\beta}_1$ under the 'just MNAR' scenario is negligible and none of the biases are statistically significant, as evident from the overlapping confidence intervals of the $\boldsymbol{\beta}_1$ estimates and the respective 'just MNAR' estimates $\overline{\boldsymbol{\beta}}^{MNAR}_{0+1,\delta_{tp}}$ and $\overline{\boldsymbol{\beta}}^{MNAR}_{0,\delta_{tp}}$ in Figures 1a and 1b. This finding indicates that a statistically significant bias in $\boldsymbol{\beta}_1$ requires more extreme MNAR scenarios, i.e. scenarios which depart much further from the MAR condition than 'just MNAR'.

Figure 2: Tipping point search for $\delta_{tp,S}$ and the corresponding $\delta_{tp,L}$ via (3')



		1981	1991	2001	2011	2015
PSID	$\delta_{tp,S}$	0.95	0.96	0.96	0.97	0.97
	$exp\ (\alpha)$	1.288	1.255	1.348	1.26	1.315
	conf. int.	(1.029, 1.613)	(1.020, 1.544)	(1.033, 1.760)	(1.001, 1.586)	(1.020, 1.695)
CPS	$\delta_{tp,L}$	0.957	0.968	0.962	0.974	0.967
	$exp\ (\alpha)$	1.288	1.256	1.343	1.259	1.322
	conf. int.	(1.196, 1.388)	(1.151, 1.370)	(1.225, 1.473)	(1.171, 1.354)	(1.228, 1.424)

Note: The figure plots the search for the tipping point using the PSID sets. Range of values for δ_{tp} [0.92; 0.97] on the x-axis of the graph with respective estimates of the odds ratio, i.e. $exp(\hat{\alpha})$ on the y-axis for different waves of the PSID sets. Tipping point is reached when $exp(\hat{\alpha}) = 1$ cannot be rejected at the 5% significance level. conf. int. denotes 95% confidence interval.

Let us examine the two more drastic MNAR scenarios: 'minimum wage MNAR' with δ_{mw} and 'husband wage MNAR' with δ_{hw}. The variety of δ values that emerge over the

decades under these two MNAR scenarios shows how SB in $\boldsymbol{\beta}_1$ varies with changing δ values; see Table 7 and the confidence interval plots of $\overline{\boldsymbol{\beta}}_{0+1,\delta_{mw}}^{MNAR}$ and $\overline{\boldsymbol{\beta}}_{0,\delta_{mw}}^{MNAR}$ versus $\overline{\boldsymbol{\beta}}_{0+1,\delta_{hw}}^{MNAR}$ and $\overline{\boldsymbol{\beta}}_{0,\delta_{hw}}^{MNAR}$ in Figures 1a and 1b. It is noticeable from the odds ratios in Table 7 that effect size magnitudes in some of these scenarios have reached a very high level, e.g. Chen *et al* (2010). Unsurprisingly, the severity of SB in $\boldsymbol{\beta}_1$ raises with MNAR cases departing further away from the MAR condition. However, the bias remains statistically insignificant as reflected by the overlapping confidence intervals of the majority of $\overline{\boldsymbol{\beta}}_{0+1,\delta_{mw}}^{MNAR}$ and $\overline{\boldsymbol{\beta}}_{0+1,\delta_{hw}}^{MNAR}$ estimates with $\boldsymbol{\beta}_1$, even in cases where the departures of the MNAR scenarios from MAR result in large differences in terms of effect size in (3). This further strengthens our finding from the first scenario of 'just MNAR' that the departure from MAR has to be relatively large to invoke even a small SB in $\boldsymbol{\beta}_1$.

Table 7. Odds Ratios From (3') with $w_{0+1,\delta}^{MNAR}$ Under δ_{mw} And δ_{hw} Scenarios

		Minimum Wage Scenario with δ_{mw}		Husband Wage Scenario with δ_{hw}	
		PSID	CPS	PSID	CPS
1981	δ	0.9	0.9	1.34 [0.0876]	1.34 [0.0444]
	Odds ratio	1.66	1.85	0.23	0.21
	(95% Conf int)	(1.328, 2.068)	(1.713, 2.008)	(0.166, 0.323)	(0.185, 0.228)
1991	δ	0.8	0.8	1.21 [0.050]	1.19 [0.0264]
	Odds ratio	3.38	4.74	0.31	0.29
	(95% Conf int)	(2.741, 4.162)	(4.288, 5.230)	(0.238, 0.401)	(0.259, 0.324)
2001	δ	0.7	0.7	1.16 [0.0272]	1.17 [0.0222]
	Odds ratio	14.83	19.22	0.27	0.30
	(95% Conf int)	(10.411, 21.123)	(16.445, 22.453)	(0.197, 0.377)	(0.266, 0.333)
2011	δ	0.6	0.6	1.11 [0.0139]	1.11 [0.0138]
	Odds ratio	37.44	105.76	0.45	0.41
	(95% Conf int)	(25.951, 54.006)	(90.826, 123.144)	(0.345, 0.594)	(0.378, 0.449)
2015	δ	0.56	0.56	1.1 [0.0123]	1.1 [0.0125]
	Odds ratio	89.89	153.34	0.44	0.46
	(95% Conf int)	(56.529, 142.927)	(127.096, 185.003)	(0.320, 0.594)	(0.424, 0.495)

Notes: Odds rations are obtained as exp ($\hat{\alpha}$) from estimating (3') by logistic regression with different MNAR simulated wage rates: $w_{0+1,\delta_{mw}}^{MNAR}$ and $w_{0+1,\delta_{hw}}^{MNAR}$. The statistics in the squared brackets for the 'husband wage' scenario are standard deviations of the ratios of $w_{0,\delta_{hw}}^{MNAR}$ to w_0^{MAR}, see equation (7), where $w_{0,\delta_{hw}}^{MNAR}$ are produced by adding a randomised mean-shift to w_0^{MAR}.

However, there are a few cases where the overlap disappears between $\boldsymbol{\beta}_1$ and $\overline{\boldsymbol{\beta}}_{0+1}^{MNAR}$, or between $\overline{\boldsymbol{\beta}}_{0+1}^{MNAR}$ and $\overline{\boldsymbol{\beta}}_{0}^{MNAR}$ under the two extreme MNAR scenarios, e.g. the PSID cases of the education variable in 2001 and 2011 and the CPS case of the age variables in 2015 under the minimum wage scenario. Given the severity of departure from MAR in these cases, it is questionable whether it is still appropriate to treat the difference as SB in $\boldsymbol{\beta}_1$, but to allow for different sub-sample coefficients. Making such a judgement requires information on the plausible MNAR mechanisms and a principled approach to handling the missing data to

identify how much and in what ways these plausible MNAR mechanisms may affect the covariates of interest.

Finally, let us compare various $\boldsymbol{\beta}_1$ with $\boldsymbol{\beta}_{1H}$. It is unsurprising to find from Figures 1a and 1b that most of $\boldsymbol{\beta}_{1H}$ do not exhibit significant deviations from $\boldsymbol{\beta}_1$, especially so when the residual correlations are small and the inverse Mill's ratio are insignificant or marginally significant, e.g. 1981 and 1991; see Table 8. However, it is puzzling why this ratio is insignificant from the PSID cases in 1981 and 1991.[8] Nevertheless, the facts that in most cases the inverse Mill's ratio is significant and that the resulting SB correction lacks significance corroborate what has been summarised in Van der Klaauw (2014). Our simulations offer an empirically trackable window to illustrate how insensitive SB can be with the MNAR condition further deviating from the MAR condition. Interestingly, where bias correction by the Hackman Procedure results in $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_{1H}$, coefficient estimates tend to coincide with $\overline{\boldsymbol{\beta}}^{MNAR}_{0+1,\delta_{mw}}$, e.g. in 2011 and 2015. It is however difficult to generalise this finding in view of the outlier case shown from $\boldsymbol{\beta}_{1H}$ of the age variable of the PSID case in 2001.

Table 8. Inverse Mill's Ratio Coefficient and Residual Correlation Estimates from the 2-step Heckman Procedure Estimation of (2)

	Estimates	1981	1991	2001	2011	2015
PSID	Inverse mill's ratio	-0.043	-0.091	0.278*	0.498**	0.549**
	(s.d.)	(0.115)	(0.107)	(0.109)	(0.177)	(0.144)
	Residual correlation	-0.075	-0.153	0.444	0.736	0.826
CPS	Inverse mill's ratio	-0.051*	0.201**	0.307**	0.508**	0.442**
	(s.d.)	(0.026)	(0.036)	(0.042)	(0.042)	(0.047)
	Residual correlation	0.097	0.353	0.511	0.758	0.666

Note: The inverse Mill's ratio is derived from a version of (3) that satisfies exclusion restrictions. Model specification of (2) is aligned with model specifications of (4). * indicates significance at the 5% and ** at the 1% significance level respectively. Given low correlation between X_i and Z_i, OVB is negligible.

5. Concluding Discussion

Our SA experiments demonstrate that the MI-based approach to assessing SB is more empirically principled than the estimator-based approach. Essentially, the MI approach enables modellers to harness information from those incomplete cases in a structured way, while this is impossible under the residual correlation assumption of the estimator-based approach. Our experiments show that while missing data is known to be the result of certain selection decisions, such as the LFP decision, these likely MNAR mechanisms do not necessarily negate

[8] Extensive specification searches of the Heckman selection model have been tried and the estimated results are quite robust.

a priori inferential validity of the parameters of interest of subjective models estimated using CC data. Modellers need to conduct carefully designed *a posteriori* SA in order to decide whether those MNAR mechanisms are ignorable for their subjective purposes. Such analyses should be aimed at assessing how much the essential features of the missing data in the incomplete case subsamples match with those in the CC subsamples, and what the plausible range of uncertainty of stochastically simulated matches may reach so that the degrees of severity of possible SB could be empirically assessed. Our experiments of the missing wage rate data show not only that the SB severity is practically ignorable for modelling wage equations within small to medium ranges of simulated mismatches under various MNAR scenarios, but also that the standard conceptualisation of SB is limiting our understanding of the mechanisms that result in SB. The idea of bias correction is predicated on the demarcation of what the appropriate population should be upon which valid inferences from CC analysis is allowed to reach or be bounded. In practice, the demarcation depends on valid sample data classification. This is essentially an aggregation issue and our experiments suggest that MI-based SA offers an empirical route to investigate this classification.

Further endeavour on matching features pertinent to human capital models between the incomplete and complete cases is desired, for instance via introducing more variables into model (4). Although stochastically imputed wage rates are not verifiable by nature of the data collection, improved matching will help raise our confidence on the judgment of whether and, if yes, to what extent the missingness mechanisms is non-ignorable. Moreover, more elaborate MNAR scenarios than the mean-shift based ones should be designed and experimented with so that our knowledge on the robustness and accuracy of the resulting SA of the risk of SB due to the missing wage rates can be enhanced.

References

Beretta, Lorenzo, and Alessandro Santaniello. 2016. "Nearest Neighbor Imputation Algorithms: A Critical Evaluation." BMC Medical Informatics and Decision Making 16 (3): 74. https://doi.org/10.1186/s12911-016-0318-z.

Blau, Francine D, and Lawrence M Kahn. 2007. "Changes in the Labor Supply Behavior of Married Women: 1980–2000." Journal of Labor Economics 25: 393–438.

Breunig, Robert, and Joseph Mercante. 2010. "The Accuracy of Predicted Wages of the Non-Employed and Implications for Policy Simulations from Structural Labour Supply Models." Economic Record 86 (272): 49–70. https://doi.org/10.1111/j.1475-4932.2009.00619.x.

Carpenter, James R., and Michael G. Kenward. 2013. Multiple Imputation and Its Application. 1st ed. Chichester, West Sussex: John Wiley & Sons.

Carpenter, James R., Michael G. Kenward, and Stijn Vansteelandt. 2006. "A Comparison of Multiple Imputation and Doubly Robust Estimation for Analyses with Missing Data." Journal of the Royal Statistical Society: Series A (Statistics in Society) 169 (3): 571–84. https://doi.org/10.1111/j.1467-985X.2006.00407.x.

Chen, Henian, Patricia Cohen, and Sophie Chen. 2010. "How Big Is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies." Communications in Statistics - Simulation and Computation 39 (4): 860–64. https://doi.org/10.1080/03610911003650383.

Heckman, James. 1974. "Shadow Prices, Market Wages, and Labor Supply." Econometrica 42 (4): 679. https://doi.org/10.2307/1913937.

———. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models." In NBER Chapters, 475–92. National Bureau of Economic Research, Inc. https://ideas.repec.org/h/nbr/nberch/10491.html.

Hsu, Chiu-Hsieh, Yulei He, Yisheng Li, Qi Long, and Randall Friese. 2016. "Doubly Robust Multiple Imputation Using Kernel-Based Techniques: Doubly Robust Multiple Imputation." Biometrical Journal 58 (3): 588–606. https://doi.org/10.1002/bimj.201400256.

Kelley, Ken, and Kristopher J. Preacher. 2012. "On Effect Size." Psychological Methods 17 (2): 137–52. https://doi.org/10.1037/a0028086.

Lin, Dan-Yu, Donglin Zeng, and Zheng-Zheng Tang. 2013. "Quantitative Trait Analysis in Sequencing Studies under Trait-Dependent Sampling." Proceedings of the National Academy of Sciences of the United States of America 110 (30): 12247. https://doi.org/10.1073/pnas.1221713110.

Little, Roderick J. A. 1988. "A Test of Missing Completely at Random for Multivariate Data with Missing Values." Journal of the American Statistical Association 83 (404): 1198–1202. https://doi.org/10.1080/01621459.1988.10478722.

Little, Roderick J. A., and Donald .B. Rubin. 1987. Statistical Analysis with Missing Data. New York: John Wiley & Sons.

Liublinska, Victoria, and Donald B Rubin. 2012. "Enhanced Tipping-Point Displays." CA: American Statistical Association, Section on Survey Research Methods, , 3861–3686.

Long, Qi, Chiu-Hsieh Hsu, and Yisheng Li. 2012. "Doubly Robust Nonparametric Multiple Imputation for Ignorable Missing Data." Statistica Sinica 22 (1). https://doi.org/10.5705/ss.2010.069.

Mercante, Joseph, and Thai-Yoong Mok. 2014. "Estimation of Labour Supply in New Zealand." Treasury Working Paper Series 14/08. New Zealand Treasury. https://econpapers.repec.org/paper/nztnztwps/14_2f08.htm.

Mincer, Jacob. 1974. "Schooling, Experience, and Earnings." NBER Books. National Bureau of Economic Research, Inc. https://econpapers.repec.org/bookchap/nbrnberbk/minc74-1.htm.

Moeller, Linda L. 2002. "On the Estimation of Classical Human Capital Wage Equations with Two Independent Sources of Data on Actual Work Experience." BLS WORKING PAPER 362, Bureau of Labor Statistics , U.S. DEPARTMENT OF LABOR.

Moffitt, Robert. 1999. "New Developments in Econometric Methods for Labor Market Analysis." Handbook of Labor Economics 3 (Part A): 1367–97.

Morris, Tim P., Ian R. White, and Patrick Royston. 2014. "Tuning Multiple Imputation by Predictive Mean Matching and Local Residual Draws." BMC Medical Research Methodology 14 (1): 75. https://doi.org/10.1186/1471-2288-14-75.

Mroz, Thomas. 1987. "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions." Econometrica 55 (4): 765–99.

Murray, Jared S. 2018. "Multiple Imputation: A Review of Practical and Theoretical Findings." ArXiv:1801.04058 [Stat], January. http://arxiv.org/abs/1801.04058.

Petreski, Marjan, Nikica Mojsoska Blazevski, and Blagica Petreski. 2014. "Gender Wage Gap When Women Are Highly Inactive: Evidence from Repeated Imputations with Macedonian Data." Journal of Labor Research 35 (4): 393–411. https://doi.org/10.1007/s12122-014-9189-1.

Puhani, Patrick. 2000. "The Heckman Correction for Sample Selection and Its Critique." Journal of Economic Surveys 14 (1): 53–68. https://doi.org/10.1111/1467-6419.00104.

Reiter, Jerome P. 2017. "Discussion: Dissecting Multiple Imputation from a Multi-Phase Inference Perspective: What Happens When God's, Imputer's and Analyst's Models Are Uncongenial?" Statistica Sinica, 1578–84.

Rubin, Donald B. 1976. "Inference and Missing Data." Biometrika 63 (3): 581–92. https://doi.org/10.2307/2335739.

———. 1987. Multiple Imputation for Nonresponse in Surveys. Wiley.

Seaman, Shaun R., Jonathan W. Bartlett, and Ian R. White. 2012. "Multiple Imputation of Missing Covariates with Non-Linear Effects and Interactions: An Evaluation of Statistical Methods." BMC Medical Research Methodology 12 (1): 46. https://doi.org/10.1186/1471-2288-12-46.

Seaman, Shaun R., and Ian R. White. 2013. "Review of Inverse Probability Weighting for Dealing with Missing Data." Statistical Methods in Medical Research 22 (3): 278–95. https://doi.org/10.1177/0962280210395740.

Seaman, Shaun R., Ian R. White, Andrew J. Copas, and Leah Li. 2012. "Combining Multiple Imputation and Inverse-Probability Weighting." Biometrics 68 (1): 129–37. https://doi.org/10.1111/j.1541-0420.2011.01666.x.

Vansteelandt, Stijn, James Carpenter, and Michael G. Kenward. 2010. "Analysis of Incomplete Data Using Inverse Probability Weighting and Doubly Robust Estimators." Methodology European Journal of Research Methods for the Behavioral and Social Sciences 6 (1): 37–48. https://doi.org/10.1027/1614-2241/a000005.

Van der Klaauw, Bas. 2014. "From Micro Data to Causality: Forty Years of Empirical Labor Economics." Labour Economics 30 (October): 88–97. https://doi.org/10.1016/j.labeco.2014.06.009.

Vella, Francis. 1998. "Estimating Models with Sample Selection Bias: A Survey." The Journal of Human Resources 33 (1): 127. https://doi.org/10.2307/146317.

Xie, Xianchao, and Xiao-Li Meng. 2017. "Dissecting Multiple Imputation from a Multi-Phase Inference Perspective: What Happens When God's, Imputer's and Analyst's Models Are Uncongenial?" Statistica Sinica. https://doi.org/10.5705/ss.2014.067.

Acknowledgment

Thanks to Ian White, James Carpenter and Luca Tasciotti for their invaluable comments and suggestions during various discussions and email exchanges concerning the design and implementation of this research project.

Appendix

Samples have been drawn from two widely used USA-based data sources: The March Annual Demographic Survey of the Current Population Survey (CPS), sponsored jointly by the Bureau of Labor Statistics and the US Census Bureau, and the Panel Study of Income Dynamics (PSID) provided by the Institute for Social Research at the University of Michigan. The Center for Economic and Policy Research (CEPR) Uniform Extracts from the CPS database are used.

The following variables have been extracted and used from both data sets: wife's hourly wage rate, wife's age, wife's education, husband's wage, number of children, and a dummy that takes on 1 if there are children under the age of 6 in the household and 0 otherwise. For the PSID data, wife's experience is also used, which is not available in the CPS data source. The coding for the education variable differs between the two datasets, hence hindering comparability. The education variable retrieved from PSID distinguishes between 8 categories coded 1 to 8, while the variable from CPS distinguishes between 6 categories coded 32.5 to 45.

The choice of the years is made based on availability of data in the two data sources. The dates stand for the year the data was published, rather than collected. A sufficient number of years is used to examine possible changes in the relationship among the variables in question. After 2000, the available years in the two databases do not match, so the comparison is made between the two closest years. More precisely, the data from PSID for 2001, 2011, and 2015 are drawn from the 2000, 2010, and 2014 sample year, respectively.

The following selection criteria are applied to both CPS and PSID data sets, excluding households with:

- Non-married, divorced, widowed or separated women.
- Women under the age of 25 and over the age of 60.
- Non-working husbands (0 wage).
- Missing data on wife's or husband's education.
- Wives or husbands working more than 4000 hours annually.
- Wives or husbands earning more than $300 USD or under $1 USD at 1999 price level per hour
- Family income net of wife's income smaller than 0.
- Wives reporting positive working hours but no wage and *vice versa*.

Table A1 provides an overview of the sample characteristics of the different years for the two datasets, with respect to the sample size and the labour force participation. Table A2 provides detailed summary statistics for the households for each year for the two datasets.

Table A1. Sample characteristics for the PSID and CPS samples.

		1981	1991	2001	2011	2015
PSID	Total Sample	2,419	3,531	2,480	2,745	2,667
	Working wife subsample	1,793	2,857	2,090	2,273	2,223
	Rate of labour force participation	74%	81%	84%	83%	83%
CPS	Total Sample	22,205	19,600	16,111	23,860	22,479
	Working wife subsample	14,493	14,644	12,299	17,690	16,540
	Rate of labour force participation	65%	74%	76%	74%	73%

Table A2. Summary statistics for the PSID and CPS samples.

		1981		1991		2001		2011		2015	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
PSID	Wife's Age	37	9.78	37.9	8.58	40.5	8.84	40.9	10.34	40.5	10.07
	Wife's Education	12.4	2.3	12.5	2.9	13.3	2.61	14.1	2.40	14.3	2.37
	Wife's Experience	7.72	6.52	10.3	7.09	10.5	7.14	9.94	7.11	9.4	7
	Wife's Hourly Wage*	6.37	4.59	10.6	8.39	15.9	11.67	21.83	17.64	23.5	16.86
	Husband's Hourly Wage	10.37	6.74	15.2	11.76	23.5	21.06	29	27.96	30.1	28.16
	Number of Children	1.45	1.27	1.44	1.26	1.25	1.20	1.17	1.25	1.24	1.28
CPS	Wife's Age	39.7	10.09	39.6	9.25	41.3	9.11	42.1	9.31	42.39	9.42
	Wife's Education	39	2.87	39.8	2.84	40.3	2.91	41	2.86	41.16	2.88
	Wife's Hourly Wage*	5.67	4.46	10.2	8.01	15.8	14.45	22.2	20.37	24.52	22.06
	Husband's Hourly Wage	9.6	5.07	15.1	9.26	23.4	22.47	29.7	26.3	32.58	31.37
	Number of Children	1.3	1.30	1.26	1.25	1.23	1.25	1.31	1.21	1.32	1.22

* conditional on working. Wage rates are nominal wages rates.