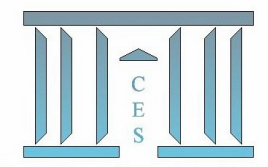
Implementing Rubin's Alternative Multiple Imputation Method for Statistical Matching in Stata

Anil Alpman

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Documents de Travail du
Centre d'Economie de la Sorbonne

CES Working Papers



Implementing Rubin's Alternative Multiple Imputation Method for Statistical Matching in Stata

Anil ALPMAN

2015.08

Maison des Sciences Économiques, 106-112 boulevard de L'Hôpital, 75647 Paris Cedex 13
http://centredeconomiesorbonne.univ-paris1.fr/
ISSN : 1955-611X

Implementing Rubin's Alternative Multiple Imputation Method for Statistical Matching in Stata

Anil Alpman *

January 29, 2015

Abstract

This paper introduces two new commands, `smpred` and `smmatch`, that implement the statistical matching procedure proposed by Rubin (1986). The purpose of statistical matching in Rubin's procedure is to generate a single dataset from various datasets, where each dataset contains a specific variable of interest and all contain some variables in common. For two variables of interest that are not observed jointly for any unit, `smpred` generates the predicted values of each as a function of the other variable of interest and a set of control variables by assuming a partial correlation value (defined by the user) between the two variables of interest (while current programs assume that they are conditionally independent given the control variables). The `smmatch` command, on the other hand, matches observations of different datasets according to their predicted values (using a minimum distance criterion) conditional on a set of control variables, and it imputes the observed value of the match for the missing.

Keywords: data combination, missing data, multiple imputation, statistical matching, `smmatch`, `smpred`.
JEL: C10, C39, C53.

1 Introduction

Statistical matching is a highly useful tool for exploring the relation between different sets of variables, for example Y and Z, which are available only in different datasets, say A and B (with the respondents in A being different than those in B, or with the impossibility to recognize the same individual that would appear in A and B because of insufficient information). If A contains Y, B contains Z, and A and B

*Paris School of Economics, University Paris 1 Panthéon-Sorbonne, CES.
E-mail: anil.alpman@univ-paris1.fr.

Documents de Travail du Centre d'Economie de la Sorbonne - 2015.08

contain common variables, for instance X, statistical matching allows to create a single dataset, say C, containing X, Y, and Z for all respondents.

This paper introduces two new commands, `smpred` and `smmatch`, that implement the statistical matching procedure proposed by Rubin (1986). Rubin's procedure is a particular kind of multiple imputation which has received much less attention than the other multiple imputation methods proposed in Rubin (1987), of which many can be applied in Stata with the use of the `mi` command. Yet, Rubin's alternative approach (i.e., Rubin [1986]) is "different from almost all other work on this topic" (Moriarity and Scheuren, 2003), and it is shown in this paper that Rubin's alternative approach yields, in the right circumstances, better results than traditional methods commonly used for statistical matching.

When Y and Z are not jointly observed for any unit in A and B, the missing values of Y are imputed most often according to the relation of Y with X, and the missing values of Z are imputed according to the relation of Z with X. Obviously, when the partial correlation value between Y and Z given X is implicitly and wrongly assumed to be zero, imputed values are likely to be inaccurate: in the most simple case, the same Y values, for example, would be imputed to units with different Z and identical X when in fact the imputed Y values should differ with Z for a given X if the partial correlation between Y and Z given X is different than zero. Moriarity and Scheuren (2003) point out that most statistical matching procedures assume, unlike Rubin (1986), that Y and Z are conditionally independent given X.

For two variables that may not be observed jointly in a dataset, the `smpred` command generates their predicted values assuming a partial correlation value (defined by the user) between these two variables given the variables in common. Said differently, given the assumed partial correlation value, `smpred` computes the regression coefficients of Y on $(1, X, Z)$ and those of Z on $(1, X, Y)$ even if Y and Z are never jointly observed. Using these regression coefficients, Y values will be predicted not only as a function of X but also as a function of Z (i.e., predicted Y values will differ with Z for a given X if the assumed partial correlation value is different than zero), and Z values will be predicted as a function of X and Y. The `smmatch` command, on the other hand, matches observations according to their predicted values (conditional on the control variables) using a minimum distance criterion and it imputes the observed value of the match for the missing.

The procedure of Rubin (1986) is summarized in Section 2. Sections 3 to 7 present the `smpred` and `smmatch` commands. Section 8 begins by providing a de-

tailed example (including practical tips and possible modifications to Rubin's original procedure) about the use of the `smpred` and `smmatch` commands, and then the results obtained with the procedure of Rubin (1986) are compared to the results obtained with two programs which are often used for statistical matching.

2 Rubin's Alternative Method for Statistical Matching

Rubin (1986) considers the situation where Y is contained in file A only, Z in file B only, and X is contained in file A and B.[1] The approach proposed by Rubin (1986) begins by a linear regression model where Y and Z are successively regressed on X:

$$Y = a_0 + aX + \epsilon \tag{1}$$

$$Z = b_0 + bX + \mu \tag{2}$$

Let α and β be the column vectors of the regression coefficients of Y on $(1, X)$ and Z on $(1, X)$, respectively. These regression coefficients may be used to generate predicted Y and Z values for the dataset formed by A and B, assuming implicitly that Y and Z are conditionally independent given X.

In many cases however, the partial correlation between Y and Z given X, denoted $\rho_{Y,Z|X}$, is different than zero. In such situations, the following matrix is constructed (Moriarity and Scheuren, 2003):

$$\begin{pmatrix} 0 & \alpha & \beta \\ -\alpha' & \text{pvar}_{Y|X} & \sigma_{Y,Z|X} \\ -\beta' & \sigma_{Y,Z|X} & \text{pvar}_{Z|X} \end{pmatrix} \tag{3}$$

where $-\alpha'$ and $-\beta'$ are the negative transposes of α and β; $\text{pvar}_{Y|X}$ is the partial variance of Y given X (which is estimated using the variances of the residuals of regression [1]); $\text{pvar}_{Z|X}$ is the partial variance of Z given X (which is estimated using the variances of the residuals of regression [2]); and $\sigma_{Y,Z|X} = \rho_{Y,Z|X}(\text{pvar}_{Y|X} * \text{pvar}_{Z|X})^{1/2}$ is the partial covariance of (Y, Z) given X. Rubin (1986) applies the sweep matrix operator to (3): sweeping on Y gives the regression coefficients of Z on $(1, X, Y)$ while sweeping on Z gives the regression coefficients of Y on $(1, X, Z)$. The new regression coefficients are used to create new predicted Y and Z values for the dataset formed by A and B.

[1] Rubin (1986) suggests to "concatenate the files and calculate a new weight for each unit" to combine the different files (e.g., A and B). Since the attribution of the new weights can be handled separately from the statistical matching procedure, its discussion is not developed in this paper.

Rubin (1986) matches each unit missing Z (i.e., units in file A) with the unit which has the closest new predicted Z value in file B, conditional on identical characteristics informed by X. Similarly, units missing Y are matched, conditional on identical characteristics informed by X, with the units which have the closest new predicted Y value in file A. Once the matches are identified, the observed value of the match is imputed for the missing value. Since $\rho_{Y,Z|X}$ is not informed by the data and uncertainty exists regarding its choice, Rubin (1986) suggests to repeat his approach assuming various values of $\rho_{Y,Z|X}$, a process resulting in multiple imputation. Therefore, multiple imputation is a helpful method to avoid erroneous conclusions according to Rubin (1986).

3 Syntax

`smpred` *depvar1 depvar2 indepvars* [*if*] [*in*] , `corr(`#`)` [`weight1(`*weightword*`)`
`weight2(`*weightword*`)` `constraint1(`*numlist*`)` `constraint2(`*numlist*`)`
`vce1(`*vcetype*`)` `vce2(`*vcetype*`)` `dropnocorr(`*dropnocorr_word*`)`]

`smmatch` *var1 var2 var3 var4 indepvars* [*if*] [*in*]

4 Description

`smpred` begins by regressing linearly *depvar1* on *indepvars*, and *depvar2* on *indepvars*. The outcome of these regressions are displayed in Stata's Results window. Using the regression coefficients obtained from these two regressions, two new variables, *pred_depvar1_0* and *pred_depvar2_0* (which correspond respectively to the predicted values of *depvar1* and *depvar2* given *indepvars*) are generated for the whole dataset. Then, using the information of these regressions and the assumed partial correlation value (defined by the `corr(#)` option) between *depvar1* and *depvar2* given *indepvars*, `smpred` computes the regression coefficients of *depvar1* on *indepvars* and *depvar2*, and of *depvar2* on *indepvars* and *depvar1* (even if *depvar1* and *depvar2* are not jointly observed). The penultimate column of the first matrix that appears in Stata's Results window reports the regression coefficients of *depvar1* on *indepvars* and *depvar2*, and the last column of the second matrix that appears in Stata's Results window reports the regression coefficients of *depvar2* on *indepvars* and *depvar1*. Using these regression coefficients, `smpred` generates new predicted values of *depvar1* and *depvar2*. These predicted values are named *pred_depvar1_#*

and *pred_depvar2_#* (or *pred_depvar1_#_m* and *pred_depvar2_#_m* if # is smaller than zero) where # corresponds to the partial correlation value between *depvar1* and *depvar2* given the *indepvars* multiplied by 100.

`smmatch` generates two new variables, *var1_imp* and *var2_imp*. For each unit missing *var1*, *var1_imp* is equal to the *var1* of the unit that has (i) an observed *var1*, (ii) a value of *var3* which is the nearest to that of the unit missing the *var1*, and (iii) the same *indepvars*. If the *var1* of the unit is observed, then *var1_imp* is equal to the unit's observed *var1*. Similarly, for each unit missing *var2*, *var2_imp* is equal to the *var2* of the unit that has (i) an observed *var2*, (ii) a value of *var4* which is the nearest to that of the unit missing the *var2*, and (iii) the same *indepvars*. If the *var2* of the unit is observed, then *var2_imp* is equal to the unit's observed var2. Note that `smmatch` supports maximum 10 *indepvars*.

If two datasets, for instance A and B, were appended, the resulting dataset must be sorted before the use of `smmatch` by a variable that allows to regroup all observations in A either before or after observations in B.

To implement the statistical matching procedure of Rubin (1986), `smmatch` must be used after `smpred` and *var1* woud correspond to *depvar1*, *var2* to *depvar2*, *var3* to *pred_depvar1_#* (or *pred_depvar1_#_m* if the assumed partial correlation value is negative), and *var4* would correspond to *pred_depvar2_#* (or *pred_depvar2_#_m* if the assumed partial correlation value is negative).

5 Options

`corr(#)` set the partaial correlation value between *depvar1* and *depvar2* given *indepvars*.

`weight1(`*weightword*`)` can be used with weighted data when *depvar1* is regressed on *indepvars*. Unlike the usual weight syntax, no brackets (i.e., [ ]) should be used. Default is `weight1()`, which means that this option is omitted.

`weight2(`*weightword*`)` can be used with weighted data when *depvar2* is regressed on *indepvars*. Unlike the usual weight syntax, no brackets (i.e., [ ]) should be used. Default is `weight2()`, which means that this option is omitted.

`constraint1(`*numlist*`)` apply specified linear constraints when *depvar1* is regressed on *indepvars*. This option can be used to suppress the constant term when regressing *depvar1* on *indepvars*. Default is `constraint1()`, which means that this option is omitted.

`constraint2(`*numlist*`)` apply specified linear constraints when *depvar2* is regressed on *indepvars*. This option can be used to suppress the constant term when regressing *depvar2* on *indepvars*. Default is `constraint2()`, which means that this option is omitted.

`vce1(`*vcetype*`)` specify the type of standard error reported when *depvar1* is regressed on *indepvars*. Default is `vce1()`, which means that this option is omitted.

`vce2(`*vcetype*`)` specify the type of standard error reported when *depvar2* is regressed on *indepvars*. Default is `vce2()`, which means that this option is omitted.

`dropnocorr(`*dropnocorr_word*`)` delete *pred_depvar1_0* and *pred_depvar2_0* (i.e., the predicted values of *depvar1* and *depvar2* when each of them are successively regressed on *indepvars*) if the expression is `yes`. Expression may be `yes` or `no`; default is `dropnocorr(no)`

6 Remarks

`smmatch` operates by creating temporary files to reduce computing time. The time required to complete the task of `smmatch` depends on the number of *indepvars* and on the number of outcomes that each *indepvar* may take. To reduce the computation time, the outcomes of variables with many outcomes (eg., age, state, or income) may be grouped within broader classes (e.g., age_group_1 may include individuals from 18 to 29, age_group_2 would include individuals from 30-39, and so forth). In addition, avoiding too few or too many indepvars may help to decrease the computation time as well.

7 Example

- `smpred weight height age male income, corr(-0.15)`
- `smpred weight height age male income, corr(0.25)`
  `weight1(aw=weight_A) weight2(aw=weight_B) constraint1(1)`
  `constraint2(3) vce1(cluster income) vce2(ro) dropnocorr(yes)`
- `smmatch weight height pred_weight_25 pred_height_25 age male`

8 Application and Comparison

To illustrate the application of the `smpred` and `smmatch` commands, and to compare them to the results obtained with two traditional methods often used for statistical matching, the first wave of the National Longitudinal Study of Adolescent Health (Add Health) is used. This dataset is collected by the Inter-university Consortium for Political and Social Research (ICPSR). In this illustration, a simple regression model, which relates weight to height, age, sex, and household income is estimated:

$$weight_i = \phi_0 + \phi_1 height_i + \phi_2 age_i + \phi_3 male_i + \phi_4 hh\_income_i + \nu_i \qquad (4)$$

where hh_income_i is the household income of individual i, ν is the error term, and ϕ_0, ϕ_1, ϕ_2, ϕ_3, and ϕ_4 are parameters to be estimated. Outliers and, for the purpose of this illustration, observations with non-informed household income are dropped. Descriptive statistics of the resulting dataset, which is referred hereafter as the *complete* dataset, are given by Table 1. The last column of Table 1 reports the regression coefficients when the model is estimated on the complete dataset. Thus, these results constitute the benchmark results over which the quality of the various statistical matching procedures explored below are evaluated.

The dataset is randomly split into two groups A and B. The weight variable is deleted in A, and the height variable is deleted in B. Thus, A and B can be con-

Table 1: Complete Dataset, Descriptive Statistics and Regression Results

Variable	Observation	Mean	Standard Deviation	Min.	Max	Regression Coefficient (standard deviation)
Weight	4769	63.91	15.60	22.68	163.29	*dependent variable*
Height	4769	1.68	0.10	1.22	2.06	78.682*** (2.232)
Age	4769	15.93	1.74	12	21	1.181*** (0.112)
Male	4769	0.50	0.50	0	1	1.180*** (0.414)
Household Income	4769	48.09	57.00	0	999	-0.014*** (0.003)

Notes: In the last column, *, **, *** indicate significance different than zero respectively at 90%, 95% and 99% confidence. Robust standard errors are in brackets.

Table 2: Incomplete Dataset, Descriptive Statistics

Variable	Observation	Mean	Standard Deviation	Min.	Max
		Group A			
Weight	0	.	.	.	.
Height	2449	1.68	0.10	1.22	2.06
Age	2449	15.92	1.75	12	21
Male	2449	0.51	0.50	0	1
Household Income	2449	48.79	57.35	0	999
		Group B			
Weight	2320	63.99	15.53	22.68	152.41
Height	0	.	.	.	.
Age	2320	15.94	1.73	12	21
Male	2320	0.48	0.50	0	1
Household Income	2320	47.35	56.63	0	999

sidered as two different datasets where each contains a different variable of interest and both contain common variables. Table 2 provides descriptive statistics of the dataset composed by A and B, which is referred hereafter as the *incomplete* dataset.

The first step of Rubin's procedure is applied with the `smpred` command which generates the predicted weight and height values for each observation of the incomplete dataset as a function of the assumed partial correlation value, denoted $\rho_{w,h|X}$, between weight and height given the control variables (i.e., age, male, and household income).[2] For example, the `smpred` command to generate the predicted weight and height values assuming that $\rho_{w,h|X} = 0.25$ is:

- `smpred weight height age male hh_income, corr(0.25) vce1(robust)`
  `> vce2(robust)`
  (Output omitted)

This command generates 4 new variables: *pred_weight_0*, *pred_height_0*, *pred_weight_25*, and *pred_height_25* (*pred_weight_25*, for example, indicates that weight is predicted by assuming that $\rho_{w,h|X} = 0.25$ while *pred_weight_0* indicates that $\rho_{w,h|X} = 0$). Table A.1 in the Appendix reports descriptive statistics of the weight and height values predicted as a function of various values of $\rho_{w,h|X}$.

An important issue, as shown in Figure 1, is the difference between the assumed and the resulting $\rho_{w,h|X}$ that may arise after the use of `smpred`. The difference between the assumed and the resulting $\rho_{w,h|X}$ may increase even more after the

[2] In this illustration, only positive partial correlation values are assumed since weight and height are positively correlated.

matching step; to prevent this increase, Moriarity and Scheuren (2003) suggest to replace the predicted values of weight and height by their observed values (when the observed value is available) before proceeding to the matching step. Since the resulting $\rho_{w,h|X}$ is a function of the assumed $\rho_{w,h|X}$, this paper suggests to construct a graph as in Figure 1 which offers a practical way to obtain the desired $\rho_{w,h|X}$ (e.g., assuming that $\rho_{w,h|X}$=0.27 will yield a resulting $\rho_{w,h|X}$=0.5). The equation of a fitted curve (e.g. a second order polynomial function) may be used to obtain more precisely the desired $\rho_{w,h|X}$.

To take into account the improvement of Moriarity and Scheuren (2003) mentioned above, the predicted values of weight in group B are replaced by their observed values and the predicted values of height are replaced by their observed values in A. Then, the procedure of Rubin (1986) is completed by using the `smmatch` command which matches each unit missing weight (i.e., units in A) with the unit in B that has, conditional on the control variables, the closest predicted value of weight; similarly, each unit missing height (i.e., units in B) is matched with the unit in A that has the closest predicted value of height conditional on the control variables. Once the

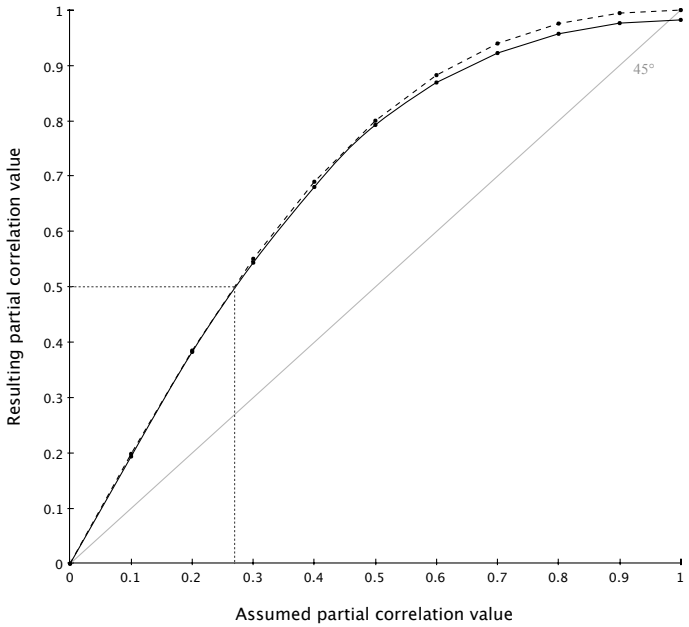


Figure 1: Assumed and resulting $\rho_{w,h|X}$ after the Use of `smpred` and `smmatch`. The dashed curve represents the relation between the assumed and the resulting $\rho_{w,h|X}$ after the use of `smpred` command. The full curve, which integrates the modification proposed by Moriarity and Scheuren (2003), displays the difference between the assumed and the resulting $\rho_{w,h|X}$ after the matching step.

Table 3: Regression Results using `smpred` and `smmatch`

Dependent Variable: Weight

Variable	`smpred` $\rho_{w,h\|X}=0$ (2.1)	`smpred` $\rho_{w,h\|X}=0.16$ (2.2)	`smpred` $\rho_{w,h\|X}=0.25$ (2.3)	`smpred` $\rho_{w,h\|X}=1$ (2.4)	`smmatch` $\rho_{w,h\|X}=0$ (3.1)	`smmatch` $\rho_{w,h\|X}=0.16$ (3.2)	`smmatch` $\rho_{w,h\|X}=0.25$ (3.3)
Height	-0.000	51.672***	78.073***	169.871***	-0.077	50.977***	76.599***
	(0.000)	(1.53)	(2.147)	(0.000)	(0.468)	(1.568)	(2.151)
Age	2.452***	1.478***	0.980***	-0.751***	2.440***	1.483***	1.008***
	(0.081)	(0.082)	(0.082)	(0.000)	(0.083)	(0.082)	(0.083)
Male	8.848***	3.488***	0.749**	-8.774***	8.831***	3.627***	0.965***
	(0.288)	(0.313)	(0.337)	(0.000)	(0.291)	(0.316)	(0.340)
Household	-0.010***	-0.013***	-0.015***	-0.021***	-0.010***	-0.013***	-0.015***
Income	(0.002)	(0.002)	(0.002)	(0.000)	(0.002)	(0.002)	(0.002)
Resulting $\rho_{Y,Z\|X}$	-0.000	0.3121	0.4707	1.0000	-0.0005	0.3107	0.4659
R^2	0.286	0.351	0.434	1.000	0.284	0.349	0.431
No. Obs.	4769	4769	4769	4769	4769	4769	4769

Notes: *, **, *** indicate significance different than zero respectively at 90%, 95% and 99% confidence. Robust standard errors are in brackets.

matches are identified, `smmatch` imputes the observed value of the match for the missing value. For example, to match observations according to predicted values generated with an assumed partial correlation value of 0.25, the `smmatch` command is

- `smmatch weight height pred_weight_25 pred_height_25 age male`

  (Output omitted)

Table A.2 in the Appendix reports descriptive statistics of the imputed weight and height values as a function of various values of $\rho_{w,h|X}$.

In Table 3, regressions (2.1) to (2.4) are performed on the predicted values (i.e., after the use of `smpred` but before the use of `smmatch`).[3] Regressions (3.1) to (3.3) are performed on the matched dataset, that is, after the use of `smmatch`. Table 3 shows that, for a given value of $\rho_{w,h|X}$, using predicted or imputed values yield similar regression coefficients. Assuming that weight and height are conditionally independent given the control variables (i.e., $\rho_{w,h|X}=0$) as in regressions (2.1) and (3.1), or over estimating the value of $\rho_{w,h|X}$ as in regression (2.4), induce biased

[3]Note that when weight or height are observed, their predicted values were replaced by their observed values.

estimates of the parameters. As the resulting value of $\rho_{w,h|X}$ gets closer to its value in the complete dataset (which is 0.47), the regression coefficients get closer to the benchmark results: Wald tests indicate that there is statistically significant difference between the benchmark results and regressions (2.3) or (3.3) only for the age variable. The results of Table 3 show therefore that Rubin's procedure produces less biased estimates than assuming conditional independence between weight and height when the value of $\rho_{w,h|X}$ is chosen within an accurate range.

Applying Rubin's statistical matching procedure with a unique partial correlation value (which is, given the purpose of statistical matching, unknown) is likely to produce inaccurate imputations if the partial correlation value is chosen incorrectly. Multiple imputation, that is, repeating Rubin's procedure with various values of the partial correlation can "help to avoid the drawing of unwarranted conclusions" (Rubin [1986]). Different imputations are combined according to two methods: first, imputations are combined by calculating the average of the imputed values following which a single regression is performed. This method can be interpreted as a pre-regression combination since the combination takes place before the regression. The second method is the combination rule proposed by Rubin (1987). Rubin's combination rule can be interpreted as a post-regression combination since the overall estimate is the average of the individual estimates and the total variance includes within and between imputation variances.

Intuitively, the partial correlation value between weight and height given the control variables is unlikely to be smaller than 0.3 and greater than 0.6. Therefore, Rubin's procedure is first repeated with assumed partial correlation values of 0.15, 0.2, 0.25, 0.3, and 0.35 which yields resulting partial correlation values of roughly 0.3, 0.4, 0.45, 0.55, and 0.6, respectively. Regression (4.1) combines these 5 partial correlation values. Given the uncertainty regarding the value of $\rho_{w,h|X}$, it might have been believed that $\rho_{w,h|X}$ could have been as small as 0.2 or greater than 0.6. Therefore, regression (4.2) includes partial correlation values of 0.1, 0.15, 0.2, 0.25, 0.3, and 0.35 (assuming that $\rho_{w,h|X} = 0.1$ yields a resulting $\rho_{w,h|X}$ of roughly 0.2); regression (4.3) includes partial correlation values of 0.15, 0.2, 0.25, 0.3, 0.35, and 0.4 (assuming that $\rho_{w,h|X} = 0.4$ yields a resulting $\rho_{w,h|X}$ of roughly 0.7); and regression (4.4) includes partial correlation values of 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, and 0.4.

When the pre-regression combination method is used, regressions (4.1) and (4.4) yield regression coefficients very close to the benchmark results: in both of these regressions, only the age coefficient is statistically different than the benchmark

Table 4: Multiple Imputation Using `smpred` and `smmatch`

	Dependent Variable: Weight			
Variable	(4.1)	(4.2)	(4.3)	(4.4)
	pre-regression combination			
Height	77.014***	70.022***	83.654***	76.916***
	(2.152)	(2.006)	(2.277)	(2.151)
Age	0.997***	1.128***	0.873***	0.999***
	(0.083)	(0.083)	(0.083)	(0.083)
Male	0.931***	1.654***	0.242	0.940***
	(0.339)	(0.333)	(0.345)	(0.339)
Household	-0.015***	-0.015***	-0.016***	-0.015***
income	(0.002)	(0.002)	(0.002)	(0.002)
R^2	0.432	0.406	0.458	0.431
No. Obs.	4769	4769	4769	4769
	Multiple imputations combined by Rubin's rule			
Height	75.756**	68.417*	81.871**	74.707*
	(23.455)	(28.381)	(26.269)	(31.215)
Age	1.020*	1.157*	0.906	1.039
	(0.530)	(0.608)	(0.572)	(0.653)
Male	1.059	1.819	0.425	1.167
	(2.490)	(2.990)	(2.776)	(3.280)
Household income	-0.015	-0.015	-0.015	-0.015
	(0.048)	(0.048)	(0.047)	(0.048)
No. Obs.	4769	4769	4769	4769

Notes: *, **, *** indicate significance different than zero respectively at 90%, 95% and 99% confidence. Robust standard errors are in brackets.

results. In regression (4.2), the height coefficient differs statistically from its benchmark coefficient but the other coefficients are statistically equal to their values in the benchmark results. In regression (4.3), only the household income coefficient is statistically equal to its coefficient in the benchmark results. Nevertheless, when compared with the results obtained using other statistical matching procedures (see below) or under the assumption that weight and height are conditionally independent given the control variables as in regression (3.1), the coefficients of height in all the regressions of Table 4 are closer to the coefficient of height estimated on the complete dataset.

Rubin's combination rule yields similar regression coefficients but higher standard deviations. Yet, the height variable remains significant. However, with Rubin's

Table 5: Regression Results using `mi` and `psmatch2` commands

Dependent Variable: Weight

Variable	`mi mvn` 5 imputations (4.1)	`mi mvn` 10 imputations (4.2)	`mi chained` 5 imputations (4.3)	`mi chained` 10 imputations (4.4)	`psmatch2` 1 neighbor (5.1)	`psmatch2` 4 neighbors (5.2)	`psmatch2` 5 neighbors (5.3)
Height	-8.54	9.56	7.63	-3.63	9.911***	11.755***	8.787***
	(28.52)	(29.45)	(14.31)	(16.26)	(2.478)	(2.081)	(1.735)
Age	2.70**	2.32***	2.33***	2.52***	2.272***	2.123***	2.066***
	(0.61)	(0.64)	(0.28)	(0.30)	(0.129)	(0.101)	(0.093)
Male	9.81**	7.94**	8.06***	9.16***	7.466***	7.678***	7.724***
	(3.00)	(3.14)	(1.74)	(1.79)	(0.484)	(0.385)	(0.356)
Household	-0.006	-0.008	-0.009	-0.008	-0.009***	-0.009***	-0.010***
income	(0.005)	(0.005)	(0.006)	(0.005)	(0.003)	(0.003)	(0.003)
No. Obs.	4769	4769	4769	4769	4769	4769	4769

Notes: *, **, *** indicate significance different than zero respectively at 90%, 95% and 99% confidence. Robust standard errors are in brackets.

combination rule, household income and male variables are not significant while age is significant only in regressions (4.1) and (4.2).

In the remainder of this section, the results obtained with the procedure of Rubin (1986) are compared to the results obtained with propensity score matching and with the multiple imputation methods proposed in Rubin (1987). The former is executed with the `psmatch2` command (Leuven and Sianesi, 2003) and the latter is performed using the `mi mvn` and the `mi chained` commands. The manual *mi impute* indicates that the `mvn` command "uses multivariate normal data augmentation to impute missing values of continuous imputation variables" while the `chained` command (i.e., a multivariate imputation using chained equations) is "another multivariate imputation method that accommodates arbitrary missing-value patterns".[4] The results obtained with these two methods are presented in Table 5.

All the regressions in Table 5 yield coefficients that are similar to those obtained under the assumption of conditional independence between weight and height given the control variables. Note however that the height coefficient is not significant in regressions (4.1) to (4.4). Unsurprisingly therefore, Wald tests indicate that all the variables in each regression of Table 5 differ statistically from the benchmark results. While the regression coefficients obtained with the propensity score matching are

[4]Since the pattern of missing values in this illustration is arbitrary, iterative methods are used. In addition, the manual *mi impute* indicates that "multiple variables usually must be imputed simultaneously ... using a multivariate imputation method".

not very sensitive to the number of nearest neighbors considered, the coefficients in regressions (4.1) to (4.4) may display some differences given the initial value of the random-number seed and the number of imputations.

9 Conclusion

This paper introduced two new commands implementing the statistical matching procedure proposed by Rubin (1986) and it showed that, for cases where the partial correlation value between the variables to be matched given the variables in common is different than zero, Rubin's procedure yields better results than matching procedures assuming implicitly that the variables to be matched are conditionally independent. Indeed, Rubin's procedure is one of the very few methods which does not assume conditional independence between the variables to be matched (Moriarity and Scheuren, 2003).

The following three points may help for a successful implementation of Rubin's procedure by preventing an incorrect assumption regarding the value of the partial correlation: first, Rubin's procedure may be repeated with various partial correlation values, an operation that amounts to multiple imputation. Then, the modification proposed by Moriarity and Scheuren (2003), which was discussed in Section 8, can be added to Rubin's original procedure. Finally, to obtain the desired partial correlation value, a graph plotting the relation between the assumed and the resulting partial correlation values can be built and, eventually, the equation of a fitted curve can be used for more precision.

Implementing Rubin's procedure through two commands (rather than one) offers two major benefits. First, modifications to Rubin's original procedure can be easily integrated as discussed in Section 8. Another possible modification may consist in using a different method to match observation after the `smpred` command (note that Rubin [1986] uses unconstrained matches). Second, both commands can be used for purposes other than statistical matching. For example, in cases where a dataset contains all the variables of interest, the `smpred` command may be used to explore how the regression coefficients would react if the partial correlation between, say, the dependant variable and one of the independent variable, was different than its value informed by the dataset.

Given the theories underlying the `mi` and the `psmatch2` commands, both have been used for statistical matching, the latter being defined as in Rubin (1986). Although Rubin's procedure yields better results in the correct context, it should

be emphasized that the `mi` command is a powerful tool for imputing "partially" missing values and analyzing the imputed values whereas `psmatch2` is a highly useful program for exploring treatment effects.

References

Leuven, E., and B. Sianesi. (2003). "PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing", http://ideas.repec.org/c/boc/bocode/s432001.html. This version 4.0.6.

Moriarity, C., and F. Scheuren. (2003). A Note on Rubin's Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. *Journal of Business & Economic Statistics*, 21, 65−73.

Rubin, D. B. (1986). Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. *Journal of Business & Economic Statistics*, 4, 87−94.

Rubin, D. B. (1987). *Multiple imputation for non-response in surveys.* New York: John Wiley & Sons.

StataCorp. (2013). "Stata Multiple-Imputation Reference Manual", College Station, TX: Stata-Corp LP.

Appendix

Table A.1: Descriptive Statistics of Predicted Values as a Function of the Assumed Partial Correlation Value.

Variable	Observation	Mean	Standard Deviation	Min.	Max
		Group A			
pred_weight_0	2449	64.17	6.27	50.04	81.40
pred_weight_10	2449	64.17	6.43	46.49	79.75
pred_weight_20	2449	64.17	6.89	42.70	82.07
pred_weight_30	2449	64.17	7.58	36.35	85.54
pred_weight_40	2449	64.17	8.47	30.01	90.78
pred_weight_50	2449	64.17	9.48	23.67	96.24
pred_weight_60	2449	64.17	10.59	17.32	101.71
pred_weight_70	2449	64.17	11.76	10.98	107.17
pred_weight_80	2449	64.17	12.98	4.63	112.64
pred_weight_90	2449	64.17	14.25	-1.71	118.11
pred_weight_100	2449	64.17	15.53	-8.05	124.13
		Group B			
pred_height_0	2320	1.68	0.06	1.56	1.84
pred_height_10	2320	1.68	0.06	1.55	1.84
pred_height_20	2320	1.68	0.06	1.54	1.87
pred_height_30	2320	1.68	0.07	1.54	1.91
pred_height_40	2320	1.68	0.07	1.53	1.96
pred_height_50	2320	1.68	0.08	1.51	2.01
pred_height_60	2320	1.68	0.08	1.50	2.05
pred_height_70	2320	1.68	0.09	1.48	2.10
pred_height_80	2320	1.68	0.09	1.47	2.15
pred_height_90	2320	1.68	0.10	1.45	2.19
pred_height_100	2320	1.68	0.10	1.44	2.24

Table A.2: Descriptive Statistics of Imputed Values as a Function of the Assumed Partial Correlation Value.

Variable	Observation	Mean	Standard Deviation	Min.	Max
			Group A		
weight_imp_0	2449	64.10	6.26	49.90	81.65
weight_imp_10	2449	64.15	6.45	47.63	81.65
weight_imp_20	2449	64.15	6.93	38.56	81.65
weight_imp_30	2449	64.15	7.60	36.29	86.18
weight_imp_40	2449	64.16	8.48	33.11	90.72
weight_imp_50	2449	64.16	9.46	31.75	95.25
weight_imp_60	2449	64.21	10.53	31.75	102.06
weight_imp_70	2449	64.23	11.61	22.68	107.50
weight_imp_80	2449	64.26	12.76	22.68	113.40
weight_imp_90	2449	64.38	13.87	22.68	120.20
weight_imp_100	2449	64.43	14.91	22.68	122.47
			Group B		
height_imp_0	2320	1.68	0.06	1.47	1.83
height_imp_10	2320	1.68	0.06	1.47	1.83
height_imp_20	2320	1.68	0.06	1.47	1.85
height_imp_30	2320	1.68	0.07	1.47	1.91
height_imp_40	2320	1.68	0.07	1.47	1.96
height_imp_50	2320	1.68	0.08	1.47	2.01
height_imp_60	2320	1.68	0.08	1.47	2.06
height_imp_70	2320	1.68	0.08	1.47	2.06
height_imp_80	2320	1.68	0.09	1.47	2.06
height_imp_90	2320	1.68	0.10	1.45	2.06
height_imp_100	2320	1.68	0.10	1.45	2.06