MatchThem:: Matching and Weighting after Multiple Imputation

by Farhad Pishgar, Noah Greifer, Clémence Leyrat and Elizabeth Stuart

Abstract Balancing the distributions of the confounders across the exposure levels in an observational study through matching or weighting is an accepted method to control for confounding due to these variables when estimating the association between an exposure and outcome and to reduce the degree of dependence on certain modeling assumptions. Despite the increasing popularity in practice, these procedures cannot be immediately applied to datasets with missing values. Multiple imputation of the missing data is a popular approach to account for missing values while preserving the number of units in the dataset and accounting for the uncertainty in the missing values. However, to the best of our knowledge, there is no comprehensive matching and weighting software that can be easily implemented with multiply imputed datasets. In this paper, we review this problem and suggest a framework to map out the matching and weighting multiply imputed datasets to 5 actions as well as the best practices to assess balance in these datasets after matching and weighting. We also illustrate these approaches using a companion package for R, **MatchThem**.

1. Introduction

In observational studies, there is the possibility that causal inferences between an exposure and an outcome may be confounded by imbalances in the distribution of the confounders across exposure groups. Balancing the distributions of these confounders across the exposure levels in the sample through matching or weighting is an accepted method to control for these confounders, to reduce the degree of dependence on certain modeling assumptions, and to obtain a less biased estimate of the causal effect.(Stuart, 2010) Despite increasing popularity in practice, these procedures cannot be immediately applied to datasets with missing values. There are several solutions to address the problem of missing data in causal effect estimation, but a standard and relatively easy-to-use one is to multiply impute the missing data, which preserves the number of units in the dataset while accounting for some of the uncertainty in the missing values.(Cham and West, 2016) However, to the best of our knowledge, there is no comprehensive matching and weighting software that facilitates causal effect estimation in multiply imputed datasets. The present paper is aimed to review the issues around matching and weighting with multiply imputed data (sections 2 and 3), to describe the steps involved in implementing best practices for these procedures (section 4), and to introduce the **MatchThem** R package (section 5), which is designed to facilitate the application of matching and weighting methods and effect estimation to multiply imputed datasets through incorporation with multiple algorithms and statistical packages.

1.1. Notation

Let $i = 1, 2, 3, ..., n$ index the n units in a dataset, in which the causal effects of a binary exposure indicator (z) on a (continuous or binary) outcome indicator (y) in the presence of a set of potential confounders ($X = \{x_1, x_2, x_3, ...\}$) are to be estimated (such that $z_i = 0$ indicates that unit i is assigned to the control group and $z_i = 1$ indicates that the unit i is assigned to the treated group) (Figure 1). Consider a situation in which the values of the some of the potential confounders or the outcome indicator are missing for a subset of units in the observed dataset. In order to account for this missingness, the missing values are multiply imputed, creating m complete datasets (such that $j = \{1, 2, 3, ..., m\}$ index these m imputed datasets). Here we focus on the procedures following imputation; see (White et al., 2011) and (Azur et al., 2011) for accessible introductions to multiple imputation for medical researchers.

1.2. Software requirements

The **MatchThem** package works with the R statistical software and programming language and can be installed within the R software (requires $\geqslant$ 3.5.0 versions) running on different platforms. **MatchThem** can be installed from the Comprehensive R Archive Network by executing the following commands in the R software console (the **MatchThem** package depends on the **MatchIt** (Ho et al., 2011) and **WeightIt** (Greifer, 2020a) packages; these lines will install those packages, too):

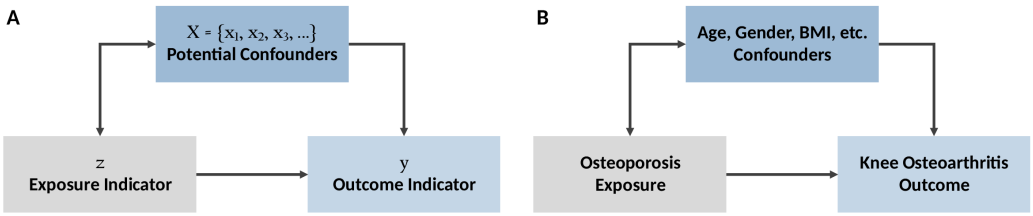

Figure 1: The Research Question. The notations used in this paper (A) and the research question used as an example in this paper (B)

```
install.packages("MatchThem")
library(MatchThem)
```

2. Matching and Weighting

Matching is a technique used to improve the robustness of the causal inferences derived from parametric and non-parametric statistical models in observational studies.(Stuart, 2010) Matching aims to control for the X (potential confounders) when estimating the relationship between z (exposure indicator) and y (outcome indicator) by duplicating, selecting, or dropping units from the dataset in a way that the resulting control and treated groups have comparable distributions for X. Despite concerns about the performance of matching methods in some instances,(King and Nielsen, 2019) if balance is achieved across the exposure groups in the matched sample, then bias in the causal effect estimate will be reduced. Typically, matching relies on a distance measure to pair similar units between exposure groups, who then form the resulting matched sample; a popular distance measure is the (different of) propensity score, the predicted probability of being assigned to the treated group given the potential confounders X. Propensity scores can be used in nearest neighbor, full, optimal, and subclassification matching (Ho et al., 2011; Williamson et al., 2012), though other distance measures and matching methods can be used as well.

Weighting is another way to achieve balance and reduce bias in a causal effect estimate. Weights for each unit can be estimated so that the distribution of potential confounders is the same across the exposure groups in the weighted samples. The weights can be used in a weighted regression of the outcome on the exposure to estimate the causal effect. A common way of estimating weights is to use a function of the propensity score, a procedure known as inverse probability weighting (IPW), though there have been some developments that bypass estimating the propensity score to estimate the weights directly.(Hainmueller, 2012; Zubizarreta, 2015)

2.1. Missing data

One of the major obstacles for most matching and weighting procedures is that they cannot be performed in a straightforward way for units with missing values in z or X because these procedures either search control and treated groups for units with exactly the same status for X or rely on the predictions from a model with z as the response variable and X as the covariates, which cannot be computed in the presence of missing data. Complete-case analysis, i.e., excluding units with missing values in the potential confounders or outcome indicator, is often the default approach for handling missing data. However, complete-case analysis may not be a valid option in all instances; the assumption of missingness completely-at-random (section 3.1), which is required to justify complete-case analysis, is often violated and it is possible that dropping units with any missing values may yield a dataset with few remaining units.(Pigott, 2001) The standard alternative to address the problem of missing data that preserves the number of units in the dataset is to multiply impute the missing values.(Leyrat et al., 2019)

3. Matching and Weighting Multiply Imputed Datasets

Given the limitations of conducting a complete-case analysis, multiply imputing missing data before applying a matching or weighting method to the dataset with missing values has become a popular alternative.

3.1. Multiply imputing missing data

Multiple imputation refers to the procedure of substituting the missing values with a set of plausible values that reflects the uncertainty in predicting the true unobserved values, which results in m imputed (filled-in) datasets.(Sterne et al., 2009) Multiple imputation is justified when the mechanism behind the missingness is ignorable, i.e., given the observed data, units with missing data represent a random subset of the dataset (*'missing-completely-at-random'* in Rubin's language (Rubin, 1987)) or when the probability that a value is missing relies on values of other observed variables, but not on the missing value itself or unobserved factors (*'missing-at-random'* in Rubin's language (Rubin, 1987)). Several multiple imputation methods are described in the literature and multiple statistical packages can be used to generate multiple imputations. Generally the broad framework of these methods is the same: impute the missing values to produce m datasets, analyze the imputed datasets separately, and pool the results obtained in each imputed dataset (using standard combining rules) to arrive at a single estimate for the sample.(Sterne et al., 2009; Rubin, 1987)

3.2. Matching and weighting multiply imputed datasets

While matching and weighting methods as the tools to estimate causal effects and multiple imputation as the flexible and general way of handling missing values are well established, there has been little work examining how to combine the two methods, and there is some debate over the correct sequence of actions for pre-processing of multiply imputed datasets by matching and weighting. There are two approaches:

1. The *within* approach: In this approach, matching or weighting is performed within each imputed dataset, using the observed and imputed covariate values, and the causal effects estimated in each of the m matched or weighted datasets are pooled together.(Leyrat et al., 2019)
2. The *across* approach: In this approach, propensity scores are averaged across the imputed datasets, and, using this averaged measure, matching or weighting is performed in the imputed datasets. Finally, the causal effects obtained from analyzing the matched or weighted datasets are pooled together.(Mitra and Reiter, 2016)

The across approach has been demonstrated to have inferior statistical performance as compared to the within approach in many common scenarios,(Leyrat et al., 2019) though early research favored its use.(Mitra and Reiter, 2016) In particular, the across approach seems most effective when outcomes are not used to impute the missing covariate values.(de Vries and Groenwold, 2016) Although some recommend avoiding the inclusion of the outcome variable during or prior to matching and weighting with propensity scores,(Rubin, 2001) statistical evidence favors the use of the outcome variable in multiple imputation of covariates.(Leyrat et al., 2019; de Vries and Groenwold, 2016) In addition, the across method is not compatible with methods that do not rely on a single distance measure for matching or weighting; such methods include (coarsened) exact matching,(de Vries and Groenwold, 2016) genetic matching,(Diamond and Sekhon, 2013) and entropy balancing,(Hainmueller, 2012) which are slowly growing in popularity due to their strong performance (It should be noted that the across approach described by Mitra and Reiter (2016) differs slightly from that described here; in their procedure, the averaged propensity scores are used to estimate the causal effect in a single dataset consisting of just the observed exposure and outcome values, which are assumed to be non-missing. The procedure described here is in the spirit of the original method but allows for the presence of imputed outcomes and the use of imputed covariates in the effect estimation. When there is no missingness in the outcome and covariates are not used in the effect estimation, the two versions of this approach coincide.).

3.3. Assessing balance in multiply imputed datasets

Balance refers to the degree to which the distribution of potential confounders is similar across the exposure groups. Typically, balance is assessed by computing the standardized mean difference (SMD) and Kolmogorov-Smirnov (KS) statistic for each covariate.(Austin and Stuart, 2015; Ho et al., 2007) When these values are small, as they would be in a randomized experiment, balance is achieved and effect estimation can proceed without fear of bias due to the observed potential confounders. Balance assessment for multiply imputed dataset has not been described previously; here we discuss best practices for balance assessment. SMDs should be computed for each covariate within each imputed dataset. Because the bias in an effect estimate is related to the mean difference of the covariates across exposure groups, and the bias in the pooled effect estimate across datasets is the average of the biases in the imputed datasets; bias can be reduced by ensuring that the average SMD for each covariate across imputed datasets is close to zero. This recommendation relies on the idea of offsetting biases: if

in some datasets the bias is positive and in others the bias is negative, on average the bias may be zero. However, even if the pooled effect estimate is unbiased, lack of balance in the individual imputed datasets can reduce the precision of the pooled estimate. Therefore, SMDs should be as close to zero as possible in each imputed dataset in addition to the average SMD across imputed datasets being small. To assess balance, we recommend the following steps:

1. Compute the SMD and KS statistic for each covariate within each imputed dataset;
2. Compute the average of the SMDs across imputed datasets; ideally, this value should be close to zero for each covariate; and
3. Compute the average and maximum of the absolute SMDs for each covariate across imputed datasets. Do the same for the KS statistics. Ideally, these should be close to zero as well, though slight departures from zero may be acceptable if the values in step 2 are close to zero.

As in datasets without missing values, the extent of balance should be assessed on interactions between covariates and their squares and cubes.(Belitser et al., 2011) In addition, the balance should be reported to ensure transparency of the analysis and to justify the validity of the estimated effect to readers. This can be done using a table or a plot, such as a *Love plot*,(Greifer, 2020b) which summarizes this information in a visually appealing and intuitive way. All of these steps can be performed by the **cobalt** R package (Greifer, 2020b), which interfaces directly with **MatchThem**.

4. Suggested Workflow

The **MatchThem** R package provides several tools and functions for proper and feasible adoption of both the within and across approaches to matching and weighting with multiply imputed data. The suggested workflow for pre-processing imputed datasets with matching or weighting using the **MatchThem** R package is as follows (Figure 2):

1. **Imputing the Missing Data in the Dataset**: There are several multiple imputation methods and statistical packages for this step. Currently, the **MatchThem** package supports imputed datasets generated by the **mice** and **Amelia** packages for R.(van Buuren and Groothuis-Oudshoorn, 2011; Honaker et al., 2011)
2. **Matching or Weighting the Imputed Datasets**: The **MatchThem** package includes functions for matching (`matchthem()`) and weighting (`weightthem()`) the multiply imputed datasets using either the within or across approaches.
3. **Assessing Balance on the Matched or Weighted Datasets**: Use functions in the **cobalt** R package to assess balance to ensure that the resulting bias is small across imputed datasets.(Greifer, 2020b) The `bal.tab()` and `love.plot()` functions in the **cobalt** package can be used directly on the output of `matchthem()` and `weightthem()`. If balance is not achieved, step 2 should be repeated with different approaches or methods until it is.
4. **Analyzing the Matched or Weighted Datasets**: Using the `with()` function from the **MatchThem** package, causal effects and their standard errors can be estimated in each matched or weighted imputed datasets. Robust standard errors should be used with weighting and most matching methods and are available through integration with the **survey** (Lumley, 2004) package.
5. **Pooling the Causal Effect Estimates**: The `pool()` function from the package can be used to pool the obtained causal effect estimates and standard errors from each dataset using Rubin's rules.

5. Example

In this section, we review the suggested workflow for matching and weighting multiply imputed datasets, using an example. The research question in this context is whether osteoporosis at baseline is associated with increased odds of developing knee osteoarthritis in the follow-up or not (Figure 1). We will use the `osteoarthritis` dataset (included in the **MatchThem** package):

```
library(MatchThem)
data("osteoarthritis")
```

The `osteoarthritis` dataset contains data on 7 characteristics (age: `AGE`, gender: `SEX`, body mass index: `BMI`, racial background: `RAC`, smoking status: `SMK`, osteoporosis at baseline: `OSP`, and knee osteoarthritis in the follow-up: `KOA`) of 2,585 individuals. The dataset contains missing data in `BMI`, `RAC`, `SMK`, `OSP`, and `KOA` variables. We assume the missing values are missing at random.

```
summary(osteoarthritis)
```

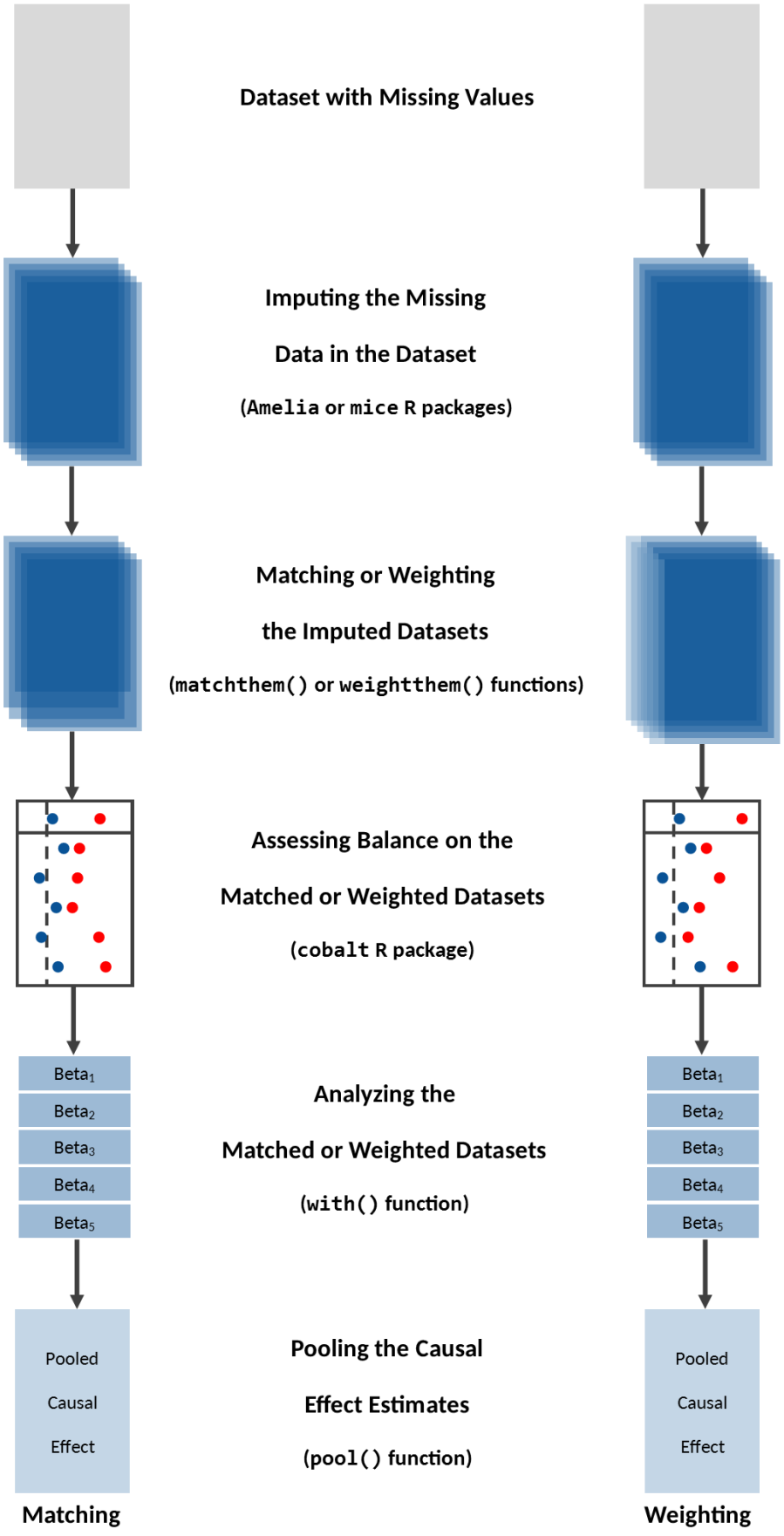

Figure 2: Suggested Workflow for Matching and Weighting Multiply Imputed Datasets

5.1. Imputing the missing data in the dataset

We use the **mice** R package to impute the missing data in the `osteoarthritis` dataset (please see the **mice** package reference manual for more details about this step (van Buuren and Groothuis-Oudshoorn, 2011)):

```
library(mice)
imputed.datasets <- mice(osteoarthritis,
                         m = 5, maxit = 10,
                         method = c("", "", "mean", "polyreg",
                                    "logreg", "logreg", "logreg"))
```

This command will produce 5 imputed datasets and save them in the `imputed.datasets` (`"mids"` class) object (the **MatchThem** package also supports imputed datasets by the **Amelia** package, please see **Amelia** package reference manual for more details (Honaker et al., 2011)).

5.2. Matching or weighting the imputed datasets

5.2.1. Matching the imputed datasets

`matchthem()` can be used to apply the within and across matching approaches and several common matching methods, including the nearest neighbor (`"nearest"`), full (`"full"`), sub-classification (`"subclass"`), optimal (`"optimal"`), exact (`"exact"`), coarsened exact (`"cem"`), and genetic (`"genetic"`) matching methods, to multiply imputed datasets (please note that only the `"nearest"`, `"full"`, `"subclass"`, and `"optimal"` matching methods are compatible with the across matching approach because other methods do not involve estimating a distance score).

In this example, we use this function to match the multiply imputed datasets, `imputed.datasets`, using all the covariates as theoretical confounders, the within matching approach, the nearest neighbor matching on the propensity score, a caliper of 5%, and the 1:2 ratio for matching (please see the package reference manual for more details):

```
matched.datasets <- matchthem(OSP ~ AGE + SEX + BMI + RAC + SMK,
                              imputed.datasets,
                              approach = 'within',
                              method = 'nearest',
                              caliper = 0.05,
                              ratio = 2)

# Matching Observations  | dataset: #1 #2 #3 #4 #5
```

After 5 iterations, the matched datasets will be produced and saved in the `matched.datasets` object (`"mimids"` class). The `"mimids"` objects contain data of the matching procedure and the matched datasets, which can be reviewed with `summary()` and `plot()` methods (e.g. `plot(matched.datasets,n = 2)`, where `n` indicates the matched dataset number), which function as they do in **MatchIt**.(Ho et al., 2011)

5.2.2. Weighting the imputed datasets

`weightthem()` can be used to estimate weights of each unit using several common weighting methods, including IPW, generalized boosted modeling weights, covariate balancing propensity score weights, and entropy balancing, in multiply imputed datasets (please see the **WeightIt** package reference manual for more details (Greifer, 2020a)).

In this example, we use this function to weight the imputed datasets, `imputed.datasets`, using all the covariates as theoretical confounders, the across weighting approach, the IPW method using logistic regression propensity scores, and targeting the average treatment effect in the matched sample (ATM) estimand (which mimics the target population resulting from matching with a caliper,(Li and Greene, 2013) please note that only methods that estimate a propensity score, which include the `"ps"`, `"gbm"`, `"cbps"`, and `"super"` weighting methods, are compatible with the across approach):

```
weighted.datasets <- weightthem(OSP ~ AGE + SEX + BMI + RAC + SMK,
                                imputed.datasets,
                                approach = 'across',
                                method = 'ps',
                                estimand = 'ATM')
```

```
# Estimating distances   | dataset: #1 #2 #3 #4 #5
# Estimating weights     | dataset: #1 #2 #3 #4 #5
```

The `weighted.datasets` object (`"wimids"` class) contains data of the weighting procedure and the weighted datasets. The `"wimids"` class objects can be reviewed with `summary()` command (e.g. `summary(weighted.datasets,n = 3)`, where `n` indicates the matched dataset number). Please note that, as in `matchthem()`, setting the `approach = 'across'`, results in adopting a slightly different across approach from the one described by Mitra and Reiter (see details in section 3.2).(Mitra and Reiter, 2016)

5.3. Assessing balance on the matched or weighted datasets

5.3.1. Assessing balance on the matched datasets

Functions of the **cobalt** package are compatible with `"mimids"` objects and the extent of the balance in the matched datasets of these objects can be checked with the `bal.tab()`, `bal.plot()`, and `love.plot()` commands:(Greifer, 2020b)

```
library(cobalt)
bal.tab(matched.datasets)
```

```
# Balance summary across all imputations
#              Type Min.Diff.Adj Mean.Diff.Adj Max.Diff.Adj
# distance Distance       0.0128        0.0138       0.0146
# AGE       Contin.      -0.0294       -0.0096       0.0154
# SEX_2      Binary      -0.0011        0.0007       0.0034
# BMI       Contin.      -0.0307       -0.0161      -0.0101
# RAC_0      Binary      -0.0011       -0.0002       0.0011
# RAC_1      Binary      -0.0078       -0.0009       0.0034
# RAC_2      Binary      -0.0045        0.0011       0.0101
# RAC_3      Binary      -0.0011        0.0000       0.0011
# SMK        Binary      -0.0157       -0.0007       0.0158
```

This information shows that the covariates (confounders) are well balanced in the osteoporosis negative and positive groups as the averaged estimated SMD for all covariates across the imputed datasets are close to zero (step 2 in section 3.3). We then assess the average and maximum of the absolute SMDs for each covariate across imputed datasets:

```
bal.tab(matched.datasets, abs = TRUE)
```

```
# Balance summary across all imputations
#              Type Mean.Diff.Adj Max.Diff.Adj
# distance Distance        0.0138       0.0146
# AGE       Contin.        0.0186       0.0294
# SEX_2      Binary        0.0016       0.0034
# BMI       Contin.        0.0161       0.0307
# RAC_0      Binary        0.0007       0.0011
# RAC_1      Binary        0.0036       0.0078
# RAC_2      Binary        0.0038       0.0101
# RAC_3      Binary        0.0004       0.0011
# SMK        Binary        0.0106       0.0158
```

The estimated average and maximum of the absolute SMDs for covariates are close to zero, meaning that the covariates are well-balanced in the imputed datasets (step 3 in section 3.3).

5.3.2. Assessing balance on the weighted datasets

The **cobalt** package is also compatible with the `"wimids"` objects and `bal.tab()`, `bal.plot()`, and `love.plot()` commands can be used on these object to assess the extent of balance the datasets of the `"wimids"` objects after weighting:(Greifer, 2020b)

```
library(cobalt)
bal.tab(weighted.datasets)
bal.tab(weighted.datasets, abs = TRUE)
```

5.4. Analyzing the matched or weighted datasets

5.4.1. Analyzing the matched datasets

The causal effect within each imputed dataset can be estimated using the `with()` command (`with()` is compatible with calls to `glm()` or similar functions with the data and weights arguments unspecified or a call to `svyglm()` or `svycoxph()` from the **survey** package with the design argument unspecified:

```
library(survey)
matched.models <- with(data = matched.datasets,
                       expr = svyglm(KOA ~ OSP, family = binomial))
```

The calculated causal effect in each matched dataset is saved in the `matched.models` object (`"mimira"` class). Please note that analyzing datasets matched with replacement, with ratios other than 1:1, or with calipers, as well as weighted datasets, requires estimating robust standard errors, which is done with `svyglm()` or `svycoxph()` from the **survey** package. In this example, we used ratio and caliper matching, hence, we adopt the robust method for estimating the standard errors.

5.4.2. Analyzing the weighted datasets

The weighted datasets can be analyzed similarly to the methods mentioned above:

```
library(survey)
weighted.models <- with(data = weighted.datasets,
                        expr = svyglm(KOA ~ OSP, family = binomial))
```

Results are saved in the `weighted.models` object (`"mimira"` class). Please note, as mentioned above, that there is no need to specify weights of units in the `expr` argument. When used with `"mimira"` class objects, the `with()` function automatically identifies the sampling or propensity score weight of each unit and performs weighted analyses.

5.5. Pooling the causal effect estimates

5.5.1. Pooling the causal effect estimates (obtained from the matched datasets)

The causal effect estimates can be pooled using the `pool()` function:

```
matched.results <- pool(matched.models)
```

The output of the `pool()` is saved in the `matched.results` object (`"mimipo"` class) and has method for `summary()` command:

```
summary(matched.results, conf.int = TRUE)

#               estimate  std.error statistic       df   p.value      2.5 %      97.5 %
# (Intercept) -0.1757256 0.09005394 -1.951337  42.5096 0.0576270 -0.3573973 0.005946089
# OSP1        -0.1580366 0.13364286 -1.182529 129.6850 0.2391593 -0.4224391 0.106365845
```

The reported result here shows that, our analysis did not find an association between the osteoporosis and knee osteoarthritis development in the follow-up in this sample (beta = -0.16 [-0.42 - 0.11], odds ratio = 0.85 [0.66 - 1.11]).

5.5.2. Pooling the causal effect estimates (obtained from the weighted datasets)

The causal effect estimates obtained from analyzing the weighted datasets can be pooled similar to the above method, using the `pool()` function:

```
weighted.results <- pool(weighted.models)
summary(weighted.results, conf.int = TRUE)

#               estimate  std.error statistic       df   p.value      2.5 %      97.5 %
# (Intercept) -0.1696747 0.07184652 -2.361628 197.1289 0.0191699 -0.3113612 -0.02798831
# OSP1        -0.1596501 0.12408788 -1.286589 222.9289 0.1995721 -0.4041854  0.08488525
```

This confirms that our analysis did not show an association between osteoporosis and knee osteoarthritis development in this sample.

6. Summary

Matching or weighting are accepted methods to balance the distributions of the confounders across the exposure levels in an observational study. However, these procedures cannot be immediately applied to datasets with missing values. Multiple imputation of the missing data is a popular approach to account for missing values while preserving the number of units in the dataset and accounting for the uncertainty in the missing values. In this paper, we suggested a framework to map out the matching and weighting multiply imputed datasets to 5 actions as well as the best practices to assess balance in these datasets after matching and weighting.

Bibliography

P. C. Austin and E. A. Stuart. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, 34(28):3661–3679, 2015. URL https://doi.org/10.1002/sim.6607. [p3]

M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf. Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1):40–49, 2011. URL https://doi.org/10.1002/mpr.329. [p1]

S. V. Belitser, E. P. Martens, W. R. Pestman, R. H. Groenwold, A. D. Boer, and O. H. Klungel. Measuring balance and model selection in propensity score methods. *Pharmacoepidemiology and Drug Safety*, 20 (11):1115–1129, 2011. URL https://doi.org/10.1002/pds.2188. [p4]

H. Cham and S. G. West. Propensity score analysis with missing data. *Psychological Methods*, 21(3): 427–445, 2016. URL https://doi.org/10.1037/met0000076. [p1]

B. P. de Vries and R. Groenwold. Comments on propensity score matching following multiple imputation. 25(6):3066—-3068, 2016. URL https://doi.org/10.1177/0962280216674296. [p3]

A. Diamond and J. S. Sekhon. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3):932–945, 2013. URL https://doi.org/10.1162/REST_a_00318. [p3]

N. Greifer. *WeightIt: Weighting for Covariate Balance in Observational Studies*, 2020a. URL https://CRAN.R-project.org/package=WeightIt. R package version 0.9.0. [p1, 6]

N. Greifer. *cobalt: Covariate Balance Tables and Plots*, 2020b. URL https://CRAN.R-project.org/package=cobalt. R package version 4.2.2. [p4, 7]

J. Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012. URL https://doi.org/10.1093/pan/mpr025. [p2, 3]

D. E. Ho, K. Imai, G. King, and E. A. Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236, 2007. URL https://doi.org/10.1093/pan/mpl013. [p3]

D. E. Ho, K. Imai, G. King, and E. A. Stuart. MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8):1–28, 2011. URL https://doi.org/10.18637/jss.v042.i08. [p1, 2, 6]

J. Honaker, G. King, and M. Blackwell. Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7):1–47, 2011. URL https://doi.org/10.18637/jss.v045.i07. [p4, 6]

G. King and R. Nielsen. Why propensity scores should not be used for matching. *Political Analysis*, 27 (4):435–454, 2019. URL https://doi.org/10.1017/pan.2019.11. [p2]

C. Leyrat, S. R. Seaman, I. R. White, I. Douglas, L. Smeeth, J. Kim, M. Resche-Rigon, J. R. Carpenter, and E. J. Williamson. Propensity score analysis with partially observed covariates: How should multiple imputation be used? *Statistical Methods in Medical Research*, 28(1):3–19, 2019. URL https://doi.org/10.1177/0962280217713032. [p2, 3]

L. Li and T. Greene. A weighting analogue to pair matching in propensity score analysis. *The International Journal of Biostatistics*, 9(2):215–234, 2013. URL https://doi.org/10.1515/ijb-2012-0030. [p6]

T. Lumley. Analysis of complex survey samples. *Journal of Statistical Software, Articles*, 9(8):1–19, 2004. URL https://doi.org/10.18637/jss.v009.i08. [p4]

R. Mitra and J. P. Reiter. A comparison of two methods of estimating propensity scores after multiple imputation. *Statistical Methods in Medical Research*, 25(1):188–204, 2016. URL https://doi.org/10.1177/0962280212445945. [p3, 7]

T. D. Pigott. A review of methods for missing data. *Educational Research and Evaluation*, 7(4):353–383, 2001. URL https://doi.org/10.1076/edre.7.4.353.8937. [p2]

D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*, volume 81. John Wiley & Sons, 1987. [p3]

D. B. Rubin. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3-4):169–188, 2001. URL https://doi.org/10.1023/A:1020363010465. [p3]

J. A. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ*, 338, 2009. URL https://doi.org/10.1136/bmj.b2393. [p3]

E. A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1–21, 2010. URL https://doi.org/10.1214/09-STS313. [p1, 2]

S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011. URL https://doi.org/10.18637/jss.v045.i03. [p4, 6]

I. R. White, P. Royston, and A. M. Wood. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399, 2011. URL https://doi.org/10.1002/sim.4067. [p1]

E. Williamson, R. Morley, A. Lucas, and J. Carpenter. Propensity scores: From naive enthusiasm to intuitive understanding. *Statistical Methods in Medical Research*, 21(3):273–293, 2012. URL https://doi.org/10.1177/0962280210394483. [p2]

J. R. Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015. URL https://doi.org/10.1080/01621459.2015.1023805. [p2]

Farhad Pishgar
Russell H. Morgan Department of Radiology and Radiological Science
Johns Hopkins University School of Medicine, Baltimore
United States
ORCiD: 0000-0003-0703-8442
Pishgar@JHMI.edu

Noah Greifer
Department of Mental Health
Johns Hopkins Bloomberg School of Public Health, Baltimore
United States
ORCiD: 0000-0003-3067-7154
NGreife1@JHU.edu

Clémence Leyrat
Department of Medical Statistics, Faculty of Epidemiology and Population Health
London School of Hygiene & Tropical Medicine, London
United Kingdom
ORCiD: 0000-0002-4097-4577
Clemence.Leyrat@lshtm.ac.uk

Elizabeth Stuart
Department of Mental Health & Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health, Baltimore
United States
ORCiD: 0000-0002-9042-8611
EStuart@JHU.edu