The Doubly Robust Estimator

MLARPIS 2026 Deliverable

Introduction

Prologue

This document is a reproducible version of the research report that the author handed in for the research seminar course. The topic of the report is the application of the doubly robust estimator in statistical matching. One remark must be made. This report does not contain real results. This is still work in progress. However, the author included a figure to show what his intentions are regarding the presentation of the results. The data for this figure are completely made up, and not based on the described methodology.

Problem Statement

Both public and private organizations heavily rely on survey data, complemented with administrative and big data sources. However, in the past decades survey response rates have been declining (Luiten, Hox, and Leeuw 2020). This introduces unit nonresponse, meaning that a sampled unit does not complete any part of the survey (Biemer 2010). Within the total survey error (TSE) framework, this nonresponse potentially introduces bias and increased variance, also referred to as unit nonresponse error (Biemer 2010; Groves and Lyberg 2010). At the same time, society demands timelier statistics, for example regarding unexpected events or when the target variables are not jointly observed (Waal 2015).

To illustrate, consider a study of the relationship between educational attainment and smoking behavior. In practice, these variables may not be observed simultaneously. In this case, estimating the association between educational attainment and smoking behavior is not trivial. However, combining data sources based on common auxiliary variables would both increase the timeliness of the statistics, and address unit nonresponse by means of enhanced data efficiency use (Waal 2015). Statistical matching provides a framework for combining data sources, for which promising results have been reported in survey applications (Donatiello et al. 2022).

Statistical Matching

In statistical matching, a statistic is computed for variables that are observed in two or more separate datasets with no or small unit overlap (D'Orazio, Zio, and Scanu 2006). In the classical statistical matching setting, two samples are considered. Each sample contains a different categorical target variable. Target variable Y is observed in sample A, while Z is observed in sample B. In both samples, a common auxiliary variable X is observed, whose levels refer to unique configurations of background characteristics. The goal is to estimate the true joint distribution of Y and Z.

To do so, additional assumptions are required. A key assumption in statistical matching is the conditional independence assumption (CIA). The CIA states that the target variables are statistically independent given their common set of auxiliary variables (D'Orazio, Zio, and Scanu 2006). Under the CIA, the association between Y and Z is fully explained by X, allowing the joint distribution to be estimated based on the marginal distributions that are observed in the separate samples.

However, in practice, the CIA is often infeasible to test (D'Orazio, Zio, and Scanu 2006). The auxiliary variables may not fully explain the association between the target variables, for example when there are unobserved confounders. Consequently, violation of the CIA can lead to biased estimates.

To reduce reliance on CIA, statistical matching can be extended adopting an external sample in which both Y and a proxy variable (Y^*) are observed. This proxy is an indirect but associated measure of target variable Y containing measurement errors. To illustrate this extension, consider the example concerning the relationship between educational attainment and smoking. Additional information can be incorporated through an external sample in which educational attainment is observed together with a related proxy, such as parental educational attainment. This external sample provides information on the underlying associations involving the target variable, which can be used to improve the statistical matching. The effectiveness of this extension depends on the selectivity of the external sample (Sojka 2025).

Selectivity in the External Sample

Selectivity is a common feature of external data sources and characterizes the external sample (Brüggen, Brakel, and Krosnick 2016). To describe these mechanisms, conditional selection probabilities are used. These probabilities denote the likelihood that unit i is included in the external sample, given the target variables, auxiliary variable and the unobserved factors that may influence this likelihood (Sojka 2025).

With respect to selection mechanisms, missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) are considered (Little and Rubin 2019; Rubin 1976). MCAR refers to the situation in which the selection probabilities are independent of the target variables, auxiliary variable or unobserved factors. MAR refers to the situation in which the selection probabilities only depend on the auxiliary variable. Finally, MNAR refers to the situation that the selection probabilities additionally depend on either the target variables, the unobserved factors, or both.

Selectivity can further be characterized by the selectivity pattern and selectivity degree (Sojka 2025). The selectivity pattern describes how the selection probabilities of unit i vary across the categories of X. The selectivity degree indicates how much the inclusion probabilities differ between the categories within a specific selectivity pattern. These probabilities are scaled such that the X category with the lowest selection probability is assigned the score 1. Selection scores for the remaining categories of X are then expressed relative to the baseline category using probability ratios. Moreover, the selectivity ratio is the probability ratio involving the baseline category, and the category of X with the highest selection score.

Doubly Robust Estimator

The joint distribution of Y and Z is primarily estimated by means of a so-called doubly robust estimator (DRE). Such an estimator generally involves two assumptions, whereby the estimate is unbiased in expectation when at least one of the assumptions holds (Bang and Robins 2005). The first assumption is that the joint distribution can be correctly estimated based on the relation between the target variable Y and the proxy Y^* derived from the external sample. The second assumption is that the joint distribution can be correctly estimated from the unit overlap in samples A and B. These two estimates are combined with the model-implied estimate, $\hat{P}(Y,Z)^{DR}$, is of the joint distribution assuming CIA validity, resulting in the doubly robust estimator shown in Equation (1). Intuitively, the doubly robust estimator corrects the overlap-based estimate using information form the external sample while adjusting for potential violations of the CIA.

$$\hat{P}(Y,X)^{DR} = \hat{P}(Y,Z)^{AB} - \hat{P}(Y,Z)^{CIA} + \hat{P}(Y,Z)^{E} \tag{1}$$

If the external model $\hat{P}(Y,Z)^{E}$ is correctly specified, the CIA estimate $\hat{P}(Y,Z)^{CIA}$ coincides with the expected value of the overlap model $\hat{P}(Y,Z)^{AB}$. Consequently, these latter two terms in Equation 1 cancel each other out, reducing the DRE to the external estimate which is then theoretically unbiased. Moreover, if the overlap model is correctly specified, the DRE is unbiased because misspecifications in the external model estimate will also be reflected in the model-implied estimate such that the expectation of $E[-\hat{P}(Y,Z)^{CIA} + \hat{P}(Y,Z)^{E}] = 0$. Finally, the DRE is unbiased as well when the CIA is valid, since both the external as overlap models are then inherently correctly specified. However, if the CIA is violated in combination with misspecification of both models, the resulting estimates will be biased in expectation (Bang and Robins 2005).

Performance of the Doubly Robust Estimator

In this section, data and design characteristics that are theoretically expected to affect the performance of the DRE are discussed. These characteristics include sample size, the degree of overlap between the samples, the quality of the proxy variable, the validity of the CIA, and the selectivity of the external sample. The performance is evaluated in terms of bias and variance of the estimated joint distribution of Y and Z.

The sizes of samples A and B are expected to affect the performance of the DRE by reducing the sampling variability of the estimated distributions used in the CIA-based and overlap components. Consequently, larger sample sizes are associated with lower variance of the DRE estimate. The degree of unit overlap in samples A and B is anticipated to affect the performance by improving the accuracy of the overlap component, which reduces the bias of the DRE in expectation. The size of the external sample is expected to affect the DRE by reducing the sampling variability of the external model component. Again, larger external sample sizes are anticipated to reduce the variance of the DRE. However, selectivity in the external sample may introduce bias into the DRE in expectation.

The quality of the proxy variable Y^* is expected to affect the performance of the DRE via its impact on the external model component. Proxy quality is characterized by the strength of the association between Y and Y^* and the structure of misclassification. Balanced misclassification implies evenly distributed misclassification across incorrect categories, whereas unbalanced misclassification implies uneven misclassification (Delden, Scholtus, and Burger 2016). Balanced misclassification is anticipated to reduce bias in the external estimate in the DRE, as conditional distributions are less distorted and Y^* remains more representative of Y.

Finally, the validity of the CIA is expected to affect the performance of the DRE, as violations of the CIA may introduce bias through the CIA-based component.

Benchmark Estimators

Iterative proportional fitting (IPF) is an alternative statistical matching procedure in which an initial joint distribution, referred to as a seed matrix, is iteratively adjusted to match known marginal distributions (Lovelace et al. 2015). IPF has been shown to perform well under various simulation settings (Lovelace et al. 2015). Therefore, IPF is suitable to serve as a benchmark estimator.

In addition, an external-sample-based estimator (EXT) is considered. This estimator only relies on information from the external sample and does not use overlap information. The EXT provides a useful benchmark for assessing the added value of combining external and overlap information in the DRE.

Present study

Taken all together, this study evaluates how data and design characteristics affect the performance of statistical matching based on a proxy variable. Accordingly, the following research question is addressed: *"To what extent do sample size, unit overlap, the validity of the conditional independence assumption, and the proxy quality affect the performance of statistical matching based on a proxy variable, and how does this compare to both iterative proportional fitting and external estimators?"*.

First, it is hypothesized that the DRE performs best when the sample sizes and unit overlap are larger. Second, the DRE is expected to exhibit lower bias when the association between the target variable and its proxy is strong, and its misclassification probabilities are balanced. Third, it is hypothesized that violations of the CIA increase bias for estimators that rely on this assumption, whereas the DRE is expected to be more robust under such violations than the benchmark estimators. Finally, it is hypothesized that the performance of the EXT is strongly associated with the representativeness of the external sample.

Methods

Data Generating Mechanisms

To evaluate the research question, a Monte Carlo simulation study is performed. For each set of simulation conditions, a synthetic population was generated, from which three samples were subsequently drawn. Table 1 provides an overview of the samples considered in this study and the categorical variables observed in each sample.

Table 1: Schematic representation of the three samples of interest, indicating the categorical variables that each sample contain with their corresponding number of levels.

Sample	Y (3)	Z (3)	X (6)	Y^* (3)
A	✓		✓	
B		✓	✓	
E	✓			✓

To model proxy quality, a transition matrix is used (Burger, Delden, and Scholtus 2015; Delden, Scholtus, and Burger 2016). The diagonal elements of the transition matrix represent the probability of correct classification p, while the off-diagonal elements w_{ij} represent misclassification probabilities, where i and j denote the categories of Y and Y^* respectively. For example, w_{12} denotes the probability that an individual belonging to category Y_1 is misclassified into category Y_2^*. The classification probabilities were assumed to be independent of both X and Y. Furthermore, misclassification probabilities were constrained to sum to one conditional on Y. The generic structure of the transition matrix is shown in Table 2.

Table 2: Structure of the transition matrix used to generate the proxy variable.

	Y_1^*	Y_2^*	Y_3^*
Y_1	p	w_{12}	w_{13}
Y_2	w_{21}	p	w_{23}
Y_3	w_{31}	w_{32}	p

Simulation Conditions

To systematically evaluate the performance of the estimators under varying data-generating settings, the study is organized into distinct simulations blocks and conditions. The simulation design distinguishes between different selection mechanisms in the external sample and varies data and design characteristics within each setting.

With respect to the external sample, two separate simulation blocks were considered. In the first block, selection was assumed to be missing at random (MAR). Four selectivity patterns were implemented: linear decrease, U-shaped, step function, and extreme increase. For the linear decrease pattern, three selectivity degrees were considered, corresponding to reduced (3:1), original (6:1), reduced (3:1) and increased (12:1) selectivity ratios.

In the second block, selection was assumed to be missing not at random (MNAR). Three main effects scenarios were considered, referred to as classic increase, non-monotonic and Y-only selection. For the classic increase scenario, four interaction effects scenarios were additionally implemented: weak, moderate, strong and extreme interaction.

Within each simulation block, additional simulation conditions were varied. Table 3 provides an overview of these conditions and their corresponding levels.

Table 3: An overview of all conditions and their corresponding dimensions in the planned Monte Carlo study, regardless of the selection mechanism in the external sample.

Condition	Number of Dimensions	Dimensions
n	3	$\{1000, 10000, 100000\}$
$\frac{n_E}{n}$	3	$\{0.20, 0.50, 0.80\}$
$\frac{n_{AB}}{n}$	3	$\{0, 0.10, 0.30\}$
p	3	$\{0.30, 0.50, 0.70\}$
$w_{ij} = \frac{1-p}{K-1},\ i \neq j$	2	$\{True, False\}$
δ	3	$\{0, 0.25, 0.75\}$

Estimator Performance Evaluation

The performance of the DRE was evaluated compared to the IPF and EXT estimators, serving as benchmarks. For each unique set of simulation conditions, all three estimators were computed.

For the IPF estimator, the joint distribution of Y and Z was estimated using a seed matrix based on the overlapping units of samples A and B. If there was no unit overlap, the seed matrix was constructed as the outer product of the target marginal distributions of Y and Z obtained from samples A and B respectively. The IPF algorithm was run per level of X and then aggregated by weighting the resulting distributions using the marginal distribution of X.

Estimator performance was evaluated in terms of bias and variance over $T = 1000$ simulation runs. Bias was assessed by computing, for each cell k of the joint distribution, the deviation of the estimated cell value, $\widehat{\theta}_k^{(t)}$), from the true cell value, θ_k, in run t. The average absolute bias per cell was then calculated across the runs and subsequently averaged across all cells of the joint distribution.

Variance across the simulation runs was estimated using Welford's algorithm (Welford 1962). This approach updates the running mean and sum of squared deviations for each cell and yields a variance estimate.

To jointly assess bias and variance, the mean squared error (MSE) was computed as the sum of squared bias and variance (Biemer, 2010). The MSE values were averaged across cells of the joint distribution, after which the square root was taken to obtain the root mean squared error (RMSE).

Finally, estimator performance was compared using relative performance measures. The relative performance Δ_C compares the average the error measure (i.e. bias, variance or RMSE) of the DRE to that of a comparative estimator C (i.e. IPF or EXT). A negative Δ_C value indicates that the DRE performs better relative to the comparative estimator and vice versa.

Hypothethical Results

Figure 1 displays the RMSE values under MNAR selection mechanisms as a function of proxy association strength, unit overlap, estimator and MNAR scenario. Across all MNAR scenarios, the DRE consistently yields the lowest RMSE values, indicating superior performance relative to both the IPF and EXT estimators. The IPF estimator shows better performance than the EXT estimator. For all estimators, stronger proxy associations are associated with lower RMSE values indicating improved accuracy. In addition, increasing unit overlap size also leads to lower RMSE values. Finally, differences across MNAR scenarios indicate that more severe forms of CIA violation result in increased estimation error, especially for the EXT estimator. The DRE remains comparatively stable across scenarios.

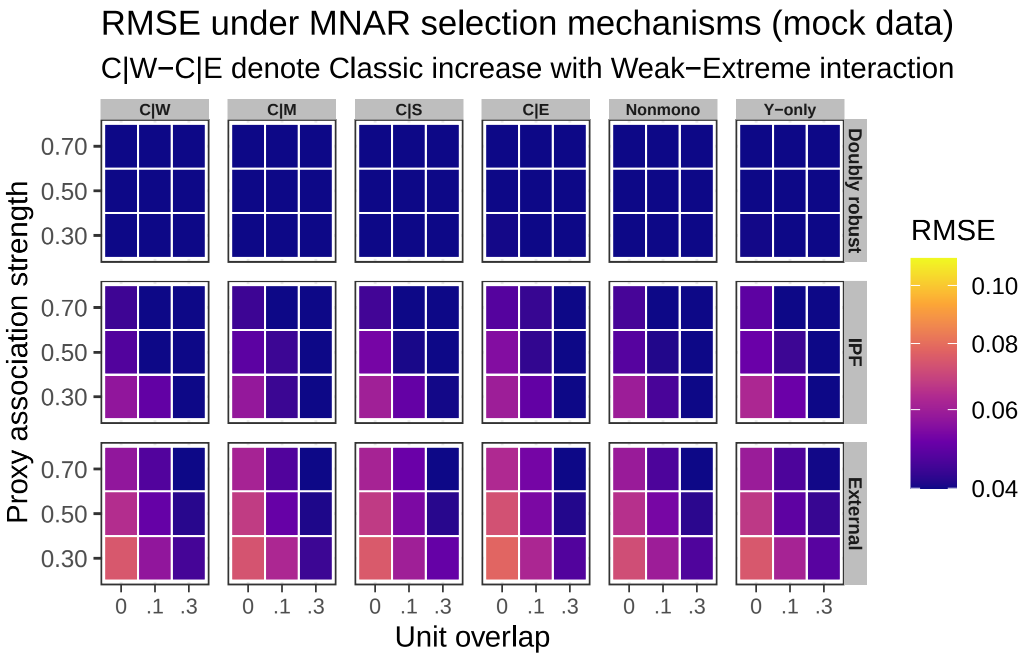

Figure 1

References

Bang, Heejung, and James M. Robins. 2005. "Doubly Robust Estimation in Missing Data and Causal Inference Models." *Biometrics* 61 (4): 962–73. https://doi.org/10.1111/j.1541-0420.2005.00377.x.

Biemer, Paul P. 2010. "Total Survey Error: Design, Implementation, and Evaluation." *Public Opinion Quarterly* 74 (5): 817–48. https://doi.org/10.1093/poq/nfq058.

Brüggen, E., J. van den Brakel, and J. Krosnick. 2016. "Establishing the Accuracy of Online Panels for Survey Research." Webpagina. *Statistics Netherlands*. https://www.cbs.nl/en-gb/background/2016/15/establishing-the-accuracy-of-online-panels-for-survey-research.

Burger, Joep, Arnout Delden, and Sander Scholtus. 2015. "Sensitivity of Mixed-Source Statistics to Classification Errors." *Journal of Official Statistics* 31 (September): 489–506. https://doi.org/10.1515/JOS-2015-0029.

D'Orazio, Marcello, Marco Di Zio, and Mauro Scanu. 2006. *Statistical Matching: Theory and Practice*. John Wiley & Sons.

Delden, Arnout van, Sander Scholtus, and Joep Burger. 2016. "Accuracy of Mixed-Source Statistics as Affected by Classification Errors." *Journal of Official Statistics* 32 (3): 619–42. https://doi.org/10.1515/jos-2016-0032.

Donatiello, Gabriella, Doriana Frattarola, Mattia Spaziani, and Marcello D'Orazio. 2022. "The Joint Distribution of Income and Consumption in Italy: An in-Depth Analysis on Statistical Matching," December. https://doi.org/10.1481/ISTATRIVISTASTATISTICAUFFICIALE_3.2022.03.

Groves, Robert M., and Lars Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly* 74 (5): 849–79. https://doi.org/10.1093/poq/nfq065.

Little, Roderick, and Donald Rubin. 2019. *Statistical Analysis with Missing Data, Third Edition*. 1st ed. Wiley Series in Probability and Statistics. Wiley. https://doi.org/10.1002/9781119482260.

Lovelace, Robin, Mark Birkin, Dimitris Ballas, and Eveline van Leeuwen. 2015. "Evaluating the Performance of Iterative Proportional Fitting for Spatial Microsimulation: New Tests for an Established Technique." *Journal of Artificial Societies and Social Simulation* 18 (2): 21.

Luiten, A, J. J. C. M. Hox, and E. D. de Leeuw. 2020. "Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study Across Countries and Surveys." *Journal of Official Statistics* 36 (3): 469–87. https://doi.org/10.2478/jos-2020-0025.

Rubin, Donald, B. 1976. "Inference and Missing Data." *Biometrika* 63 (3): 581–92. https://doi.org/10.1093/biomet/63.3.581.

Sojka, B. L. 2025. "Statistical Matching Using a Non-Probability Sample as Auxiliary Dataset."

Waal, A. G. de. 2015. *Statistical Matching: Experimental Results and Future Research Questions*. Den Haag: CBS.

Welford, B. P. 1962. "Note on a Method for Calculating Corrected Sums of Squares and Products." *Technometrics* 4 (3): 419–20. https://doi.org/10.1080/00401706.1962.10490022.