Burgard, Jan Pablo; Krause, Joscha; Münnich, Ralf T.

Working Paper

A study of discontinuity effects in regression inference based on web-augmented mixed mode surveys

Research Papers in Economics, No. 3/20

Provided in Cooperation with:
University of Trier, Department of Economics

Suggested Citation: Burgard, Jan Pablo; Krause, Joscha; Münnich, Ralf T. (2020) : A study of discontinuity effects in regression inference based on web-augmented mixed mode surveys, Research Papers in Economics, No. 3/20, Universität Trier, Fachbereich IV - Volkswirtschaftslehre, Trier

This Version is available at:
https://hdl.handle.net/10419/243470

Universität Trier

A Study of Discontinuity Effects in Regression Inference based on Web-Augmented Mixed Mode Surveys

Jan Pablo Burgard
Joscha Krause
Ralf Münnich

Research Papers in Economics
No. 3/20

A Study of Discontinuity Effects in Regression Inference based on Web-Augmented Mixed Mode Surveys

Jan Pablo Burgard, Joscha Krause, Ralf Münnich

Department of Economic and Social Statistics

Abstract

We consider a situation where the sample design of a survey is modified over time in order to save resources. The former design is a classical large-scale survey. The new design is a mixed mode survey where a smaller classical sample is augmented by records of an online survey. For the online survey no inclusion probabilities are available. We study how this change of data collection affects regression coefficient estimation when the model remains constant in the population over time. Special emphasis is placed on situations where the online records are selective with respect to the model. We develop a statistical framework to quantify so-called survey discontinuities in regression analysis. The term refers to differences between coefficient estimates that solely stem from the survey redesign. For this purpose, we apply hypothesis tests to identify whether observed differences in estimates are significant. Further, we discuss propensity estimation and calibration as potential methods to reduce selection biases stemming from the web survey. A Monte Carlo simulation study is conducted to test the methods under different degrees of selectivity. We find that even mild informativeness significantly impairs regression inference relative to the former survey despite bias correction.

Keywords: Calibration, hypothesis test, informative sampling, propensity score estimation

1 Introduction

Survey samples long have been the primary data sources for empirical analysis in various research fields, such as economics, sociology, and political science. For the collection of a survey sample, essential features like sampling frame and sample design must be defined in order to associate every individual of the target population with a non-zero inclusion probability. After the sampling process, these probabilities are anticipated in the statistical analysis to allow for sound inference. See Särndal et al. (1992) or Fuller (2009) for a comprehensive overview. However, the collection of exhaustive survey samples – especially on national levels – is typically very costly. Therefore, sample sizes are more and more reduced by policy-makers in order to save resources. Naturally, this leads to an increase in sample variance that may impair estimation quality beyond acceptable levels. In this case,

alternative data sources such as online surveys are often considered to augment classical sample observations by reducing variance in the estimation process. The combination of a classical survey and an online survey is called *web-augmented mixed mode survey*. It usually contains a comparable number of sampled individuals relative to a classical exhaustive survey, but is faster and considerably cheaper to collect.

There is an ongoing debate to what extent web-augmented mixed mode surveys represent true alternatives to classical large-scale surveys. An important drawback of online data is that it is typically not possible to quantify inclusion probabilities since the sampling frame is unknown. In some situations, there is information available that can be used to approximate the unknown inclusion probabilities, for instance via propensity score estimation (Rosenbaum and Rubin, 1983) or calibration methods (Deville and Särndal, 1992). However, in the absence of such information, the sample observations have to be treated as the result of a simple random sample. This marks a major issue for statistical inference, as online surveys are known to be affected by informative sampling due to framing errors and coverage problems (Zagheni and Weber, 2015). The term *informative sampling* refers to situations where the inclusion probabilities are not independent from the outcomes of a statistical model after conditioning on auxiliary data (Pfeffermann and Sverchkov, 2009). If not accounted for, this can lead to severe bias in the empirical analysis. Therefore, it has to be carefully evaluated how web-augmented data collection alters estimation outcomes relative to a classical survey, and whether it really improves their quality.

A suitable concept for a corresponding evaluation is called *survey discontinuity* (van den Brakel et al., 2008). It quantifies the difference between two estimates of a given statistic that solely stems from a change in the preceding sampling process. In the literature, survey discontinuities are typically assessed via model-based time series analysis. On that note, van den Brakel and Roels (2010) use a state-space intervention model for the estimation of discontinuities in a Dutch survey on social participation and environmental consciousness. Smith et al. (2017) study potential discontinuities in the National Survey for Wales. Further, van den Brakel et al. (2020) compare the state-space intervention approach with a structural time series model that is combined with a parallel run of the former survey design. These studies provide important insights into how the predictive inference based on corresponding models is affected by the redesign of a survey.

In this paper, survey discontinuities are investigated from a different perspective. On the example of linear regression, we study how inference with respect to the model parameters themselves is affected by changes in the sampling process. For this, we consider a situation where the sample design of a survey is modified over time in order to save resources. The former design is a classical large-scale survey, while the new design is a web-augmented mixed mode survey consisting of a small classical sample and records of an online survey. For latter, no information about inclusion probabilities is available. Under the assumption that the model is constant in the population over time, we develop a statistical framework to quantify sampling-related differences in regression coefficient estimates between both periods. Hypothesis tests are used to assess whether found differences are significant under the null hypothesis of equality. Further, we address the issue of informative sampling in on-

line surveys by discussing propensity score estimation and calibration as potential methods for approximating the missing inclusion probabilities. An extended Monte Carlo simulation study is conducted where the corresponding setting is implemented based on the synthetic dataset AMELIA (Burgard et al., 2017). We consider different scenarios with respect to the degrees of informativity associated with the augmenting online survey records. We find that even mild informativeness of the augmenting data significantly impairs regression inference relative to the former survey despite using correction methods.

The remainder of the paper is organized as follows. In Section 2, the statistical framework to quantify the impact of the data collection change on regression coefficient estimation as well as the correction methods are presented. Section 3 contains the simulation study as well as a critical analysis of its results. Section 4 closes with some conclusive remarks and an outlook on future research.

2 Theory

We first present the statistical framework to quantify differences between regression coefficient estimates that may indicate survey discontinuities resulting from web-augmentation. Thereafter, we present propensity score estimation and calibration as correction methods for bias stemming from missing inclusion probabilities and informative sampling as a result of self-selectivity in online records.

2.1 Sampling

For the subsequent developments, we follow the definitions for finite population inference based on survey sampling provided by Cassel et al. (1977), Chapter 1. Consider a finite population at two time periods $t \in \{1, 2\}$. Let $\mathcal{U}_1 = \{1, ..., N_1\}$ denote the population in period $t = 1$ containing $|\mathcal{U}_1| = N_1$ individuals indexed by $i = 1, ..., N_1$. Likewise, let $\mathcal{U}_2 = \{1, 2, ..., N_2\}$ be the population in period $t = 2$ containing $|\mathcal{U}_2| = N_2$ individuals indexed by $i = 1, ..., N_2$. For the first period, assume that a survey sample $\mathcal{S}_1 \subset \mathcal{U}_1$ of size $|\mathcal{S}_1| = n_1 < N_1$ is drawn from $\mathcal{U}_1$ under a given sample design. We define the term *sample design* as a function that associates every possible subset of the required size from the population of a period with a probability of being chosen. Thus, for period $t = 1$, the sample design is formally given by $P_1 : \mathbb{S}_1 \to [0, 1]$ with $\mathbb{S}_1 = \{\mathcal{S}_1 : \mathcal{S}_1 \subset \mathcal{U}_1 \wedge |\mathcal{S}_1| = n_1\}$ and $\sum_{\mathcal{S}_1 \in \mathbb{S}_1} P_1(\mathcal{S}_1) = 1$. The inclusion probability for some $i \in \mathcal{S}_1$ is denoted by

$$\pi_{1i} := \Pr(i \in \mathcal{S}_1) = \sum_{\mathcal{S}_1 \in \mathbb{S}_1} \mathbb{1}_{(i \in \mathcal{S}_1)} P_1(\mathcal{S}_1), \tag{1}$$

where $\mathbb{1}_{(\cdot)}$ is the indicator function. For the second period, let a survey sample $\mathcal{S}_2$ of size $|\mathcal{S}_2| = n_2 < N_2$ and $(n_2/N_2) < (n_1/N_1)$ be drawn from $\mathcal{U}_2$. Suppose that the sample design $P_2 : \mathbb{S}_2 \to [0, 1]$ with $\mathbb{S}_2 = \{\mathcal{S}_2 : \mathcal{S}_2 \subset \mathcal{U}_2 \wedge |\mathcal{S}_2| = n_2\}$ and $\sum_{\mathcal{S}_2 \in \mathbb{S}_2} P_2(\mathcal{S}_2) = 1$ is different from the design in the last period. The inclusion probability for some $i \in \mathcal{S}_2$ is denoted by

$$\pi_{2i} := \Pr(i \in \mathcal{S}_2) = \sum_{\mathcal{S}_2 \in \mathbb{S}_2} \mathbb{1}_{(i \in \mathcal{S}_2)} P_2(\mathcal{S}_2). \tag{2}$$

Let $\mathcal{D} \subset \mathcal{U}_2$ be an additional subset of the population in period $t = 2$ with $|\mathcal{D}| = n_{\mathcal{D}}$. The subset represents the online dataset that is used for augmenting the records from the smaller sample in $t = 2$. In light of Section 1, suppose that there is no information available on how $\mathcal{D}$ is collected. Although, for simplicity, assume that $\mathcal{D} \cap \mathcal{S}_2 = \emptyset$.

2.2 Regression Coefficient Estimation

Let y be a real-valued response variable of interest with realization $y_{it} \in \mathbb{R}$ for some $i \in \mathcal{U}_t$. Denote $x = \{x_1, ..., x_p\}$ as a set of real-valued covariates statistically related to y with realization $\mathbf{x}_{it} \in \mathbb{R}^{1\times p}$ for $i \in \mathcal{U}_t$. Suppose the relation for any $i \in \mathcal{U}_t$ is characterized by

$$y_{ti} = \mathbf{x}_{ti}\boldsymbol{\beta} + e_{ti}, \quad e_{ti} \overset{iid}{\sim} N(0, \sigma^2), \tag{3}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ is a vector of unknown regression coefficients and e_{it} is a random model error with variance parameter $\sigma^2 > 0$. Note that we assume (3) to hold for all individuals and time periods. Hence, $\boldsymbol{\beta}$ is constant for $t = 1, 2$. Let the pair $(y_{ti}, \mathbf{x}_{ti})$ be observed for all individuals in $\mathcal{S}_1$ as well as $\mathcal{S}_2 \cup \mathcal{D}$. The objective is to estimate $\boldsymbol{\beta}$ from each temporal data basis individually. For $t = 1$, this can be done via weighted least squares according to

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_1 &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i\in\mathcal{S}_1} \pi_{1i}^{-1} \left(y_{1i} - \mathbf{x}_{1i}\boldsymbol{\beta}\right)^2 \\
&= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left(\mathbf{y}_1 - \mathbf{X}_1\boldsymbol{\beta}\right)'\boldsymbol{\Pi}_1^{-1}(\mathbf{y}_1 - \mathbf{X}_1\boldsymbol{\beta}) \\
&= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \mathbf{y}_1'\boldsymbol{\Pi}_1^{-1}\mathbf{y}_1 - \mathbf{y}_1'\boldsymbol{\Pi}_1^{-1}\mathbf{X}_1\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}_1'\boldsymbol{\Pi}_1^{-1}\mathbf{y}_1 + \boldsymbol{\beta}'\mathbf{X}_1'\boldsymbol{\Pi}_1^{-1}\mathbf{X}_1\boldsymbol{\beta},
\end{aligned} \tag{4}$$

where $\mathbf{y}_1 = (y_{11}, ..., y_{1n_1})'$, $\mathbf{X}_1 = (\mathbf{x}_{11}', ..., \mathbf{x}_{1n_1}')'$, and $\boldsymbol{\Pi}_1 = \operatorname{diag}(\pi_{11}, ..., \pi_{1n_1})$. We differentiate with respect to $\boldsymbol{\beta}$ and set the gradient to zero:

$$\nabla_{\widehat{\boldsymbol{\beta}}_1} = 2\left(-\mathbf{X}_1'\boldsymbol{\Pi}_1^{-1}\mathbf{y}_1 + \mathbf{X}_1'\boldsymbol{\Pi}_1^{-1}\mathbf{X}_1\boldsymbol{\beta}\right) \overset{!}{=} \mathbf{0}_p. \tag{5}$$

Solving for $\boldsymbol{\beta}$ then yields the well-known weighted least squares estimator

$$\widehat{\boldsymbol{\beta}}_1 = \left(\mathbf{X}_1'\boldsymbol{\Pi}_1^{-1}\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\boldsymbol{\Pi}_1^{-1}\mathbf{y}_1. \tag{6}$$

Let $\mathbf{e}_1 = (e_{11}, ..., e_{1n_1})'$ denote the model error vector of all sampled individuals $i \in \mathcal{S}_1$ with $\operatorname{Var}(\mathbf{e}_1) = \sigma^2\mathbf{I}_{n_1}$, where $\mathbf{I}_{n_1}$ is the $(n_1 \times n_1)$-identity matrix. Since (6) is an unbiased estimator of $\boldsymbol{\beta}$ and $\mathbf{e}_1$ is the only random component, its variance is given by

$$\begin{aligned}
\operatorname{Var}(\widehat{\boldsymbol{\beta}}_1) &= \mathrm{E}\left[\left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\right)'\right] \\
&= \mathrm{E}\left[\left(\mathbf{X}_1'\boldsymbol{\Pi}_1^{-1}\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\boldsymbol{\Pi}_1^{-1}\mathbf{e}_1\left(\left(\mathbf{X}_1'\boldsymbol{\Pi}_1^{-1}\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\boldsymbol{\Pi}_1^{-1}\mathbf{e}_1\right)'\right] \\
&= \left(\mathbf{X}_1'\boldsymbol{\Pi}_1^{-1}\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\boldsymbol{\Pi}_1^{-1}\sigma^2\mathbf{I}_{n_1}\boldsymbol{\Pi}_1^{-1}\mathbf{X}_1\left(\mathbf{X}_1'\boldsymbol{\Pi}_1^{-1}\mathbf{X}_1\right)^{-1}.
\end{aligned} \tag{7}$$

For $t = 2$, we have to pool the observations from $\mathcal{S}_2$ and $\mathcal{D}$ to create a combined objective function for regression coefficient estimation. Recall that we have information on inclusion

probabilities for sampled individuals in $\mathcal{S}_2$, but not for individuals in $\mathcal{D}$. Under this premise, the weighted least squares estimator of $\boldsymbol{\beta}$ is the solution to the optimization problem

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_2 &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i\in\mathcal{S}_2} \pi_{2i}^{-1} \left(y_{2i} - \mathbf{x}_{2i}\boldsymbol{\beta}\right)^2 + \sum_{i\in\mathcal{D}} \frac{N_2}{n_{\mathcal{D}}} \left(y_{\mathcal{D}i} - \mathbf{x}_{\mathcal{D}i}\boldsymbol{\beta}\right)^2 \\
&= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left((\mathbf{y}_2', \mathbf{y}_{\mathcal{D}}')' - (\mathbf{X}_2', \mathbf{X}_{\mathcal{D}}')'\boldsymbol{\beta}\right)' \boldsymbol{\Pi}_{2\mathcal{D}}^{-1} \left((\mathbf{y}_2', \mathbf{y}_{\mathcal{D}}')' - (\mathbf{X}_2', \mathbf{X}_{\mathcal{D}}')'\boldsymbol{\beta}\right) \\
&= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} (\mathbf{y}_2', \mathbf{y}_{\mathcal{D}}')\boldsymbol{\Pi}_{2\mathcal{D}}^{-1}(\mathbf{y}_2', \mathbf{y}_{\mathcal{D}}')' - (\mathbf{y}_2', \mathbf{y}_{\mathcal{D}}')\boldsymbol{\Pi}_{2\mathcal{D}}^{-1}(\mathbf{X}_2', \mathbf{X}_{\mathcal{D}}')'\boldsymbol{\beta} \\
&\qquad - \boldsymbol{\beta}'(\mathbf{X}_2', \mathbf{X}_{\mathcal{D}}')\boldsymbol{\Pi}_{2\mathcal{D}}^{-1}(\mathbf{y}_2', \mathbf{y}_{\mathcal{D}}')' + \boldsymbol{\beta}'(\mathbf{X}_2', \mathbf{X}_{\mathcal{D}}')\boldsymbol{\Pi}_{2\mathcal{D}}^{-1}(\mathbf{X}_2', \mathbf{X}_{\mathcal{D}}')'\boldsymbol{\beta}
\end{aligned} \tag{8}$$

where $\mathbf{y}_2 = (y_{21}, ..., y_{2n_2})'$, $\mathbf{y}_{\mathcal{D}} = (y_{\mathcal{D}1}, ..., y_{\mathcal{D}n_{\mathcal{D}}})'$, $\mathbf{X}_2 = (\mathbf{x}_{21}', ..., \mathbf{x}_{2n_2}')'$, $\mathbf{X}_{\mathcal{D}} = (\mathbf{x}_{\mathcal{D}1}', ..., \mathbf{x}_{\mathcal{D}n_{\mathcal{D}}}')'$. Further, $\boldsymbol{\Pi}_{2\mathcal{D}}$ is a $[(n_2 + n_{\mathcal{D}}) \times (n_2 + n_{\mathcal{D}})]$-diagonal matrix where $\pi_{21}, ..., \pi_{2n_2}$ are the first n_2 main diagonal elements and the remaining $n_{\mathcal{D}}$ elements are given by $n_{\mathcal{D}}/N_2$. As before, we differentiate with respect to $\boldsymbol{\beta}$ and set the gradient to zero:

$$\nabla_{\widehat{\boldsymbol{\beta}}_2} = 2\left(-(\mathbf{X}_2', \mathbf{X}_{\mathcal{D}}')\boldsymbol{\Pi}_{2\mathcal{D}}^{-1}(\mathbf{y}_2', \mathbf{y}_{\mathcal{D}}')' + (\mathbf{X}_2', \mathbf{X}_{\mathcal{D}}')\boldsymbol{\Pi}_1^{-1}(\mathbf{X}_2', \mathbf{X}_{\mathcal{D}}')'\boldsymbol{\beta}\right) \stackrel{!}{=} \mathbf{0}_p. \tag{9}$$

Solving for $\boldsymbol{\beta}$ yields the weighted least squares estimator based on the combined data

$$\widehat{\boldsymbol{\beta}}_2 = \left((\mathbf{X}_2', \mathbf{X}_{\mathcal{D}}')\boldsymbol{\Pi}_{2\mathcal{D}}^{-1}(\mathbf{X}_2', \mathbf{X}_{\mathcal{D}}')'\right)^{-1} (\mathbf{X}_2', \mathbf{X}_{\mathcal{D}}')\boldsymbol{\Pi}_{2\mathcal{D}}^{-1}(\mathbf{y}_2', \mathbf{y}_{\mathcal{D}}')'. \tag{10}$$

Note that with this specification of $\boldsymbol{\Pi}_{2\mathcal{D}}$, we have to assume that the observations in $\mathcal{D}$ have been collected via simple random sampling, or at least such that they are non-informative with respect to the model. That is to say, $\mathrm{E}(y_i|\mathbf{x}_i) = \mathrm{E}(y_i|\mathbf{x}_i, \mathbb{1}_{(i\in\{\mathcal{S}_2\cup\mathcal{D}\})})$ must be fulfilled for all $i \in \mathcal{U}_2$. By the comments on online survey records in Section 1, this may not hold in practice. In the worst case, it leads to $\widehat{\boldsymbol{\beta}}_2$ being a biased estimator of $\boldsymbol{\beta}$. Let $\mathbf{e}_2 = (e_{21}, ..., e_{2n_2})'$ and $\mathbf{e}_{\mathcal{D}} = (e_{\mathcal{D}1}, ..., e_{\mathcal{D}n_{\mathcal{D}}})'$ denote the model error vectors for all $i \in \mathcal{S}_2$ and $i \in \mathcal{D}$, respectively, with $\mathrm{Var}(\mathbf{e}_2', \mathbf{e}_{\mathcal{D}}') = \sigma^2\mathbf{I}_{n_2+n_{\mathcal{D}}}$. Provided the non-informativity assumption is fulfilled and by the argumentation for (7), the variance of $\widehat{\boldsymbol{\beta}}_2$ is given by

$$\begin{aligned}
\mathrm{Var}(\widehat{\boldsymbol{\beta}}_2) &= \mathrm{E}\left[\left(\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}\right)'\right] \\
&= \mathrm{E}\left[\mathbf{A}^{-1}(\mathbf{X}_2', \mathbf{X}_{\mathcal{D}}')\boldsymbol{\Pi}_{2\mathcal{D}}^{-1}(\mathbf{e}_2', \mathbf{e}_{\mathcal{D}}')' \left(\mathbf{A}^{-1}(\mathbf{X}_2', \mathbf{X}_{\mathcal{D}}')\boldsymbol{\Pi}_{2\mathcal{D}}^{-1}(\mathbf{e}_2', \mathbf{e}_{\mathcal{D}}')'\right)'\right] \\
&= \mathbf{A}^{-1}(\mathbf{X}_2', \mathbf{X}_{\mathcal{D}}')\boldsymbol{\Pi}_{2\mathcal{D}}^{-1}\sigma^2\mathbf{I}_{n_2+n_{\mathcal{D}}}\boldsymbol{\Pi}_{2\mathcal{D}}^{-1}(\mathbf{X}_2', \mathbf{X}_{\mathcal{D}}')'\mathbf{A}^{-1},
\end{aligned} \tag{11}$$

where $\mathbf{A} = (\mathbf{X}_2', \mathbf{X}_{\mathcal{D}}')\boldsymbol{\Pi}_{2\mathcal{D}}^{-1}(\mathbf{X}_2', \mathbf{X}_{\mathcal{D}}')'$.

2.3 Hypothesis Test

We now present the concept of survey discontinuities in our setting. For this, we have to assess whether the regression coefficient estimates based on $\mathcal{S}_1$ as well as $\mathcal{S}_2 \cup \mathcal{D}$ differ significantly. If the model (3) holds in the population over time and the web-augmented mixed mode survey $\mathcal{S}_2 \cup \mathcal{D}$ is not affected by informative sampling, then there should be no

differences between estimates $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ on expectation. In order to evaluate this aspect quantitatively, we follow Clogg et al. (1995) and Pasternoster et al. (1998) by using a t-test for two independent samples that tests the equality of regression coefficient estimates. For two given estimates $\widehat{\beta}_{1j} \in \widehat{\boldsymbol{\beta}}_1$ and $\widehat{\beta}_{2j} \in \widehat{\boldsymbol{\beta}}_2$, the null hypothesis is $H_0 : \widehat{\beta}_{1j} = \widehat{\beta}_{2j}$. The corresponding test statistic is given by

$$T_j = \frac{\widehat{\beta}_{1j} - \widehat{\beta}_{2j}}{\sqrt{\mathrm{Var}(\widehat{\beta}_{1j}) + \mathrm{Var}(\widehat{\beta}_{2j})}}, \tag{12}$$

where $\mathrm{Var}(\widehat{\beta}_{1j})$ and $\mathrm{Var}(\widehat{\beta}_{2j})$ are the j-th elements of $\mathrm{Var}(\widehat{\boldsymbol{\beta}}_1)$ and $\mathrm{Var}(\widehat{\boldsymbol{\beta}}_2)$, respectively. The test statistic approximately follows a standard normal distribution for n_1 and $n_2 + n_{\mathcal{D}}$ sufficiently large. A deviation between estimates is called *significant* if we have

$$T_j \notin [z(\alpha/2); z(1-\alpha/2)], \quad \alpha \in (0,1) \tag{13}$$

for some significance level α and the related quantile $z(\cdot)$ of the standard normal distribution. Please note that even if $\mathrm{E}(\widehat{\boldsymbol{\beta}}_1) = \mathrm{E}(\widehat{\boldsymbol{\beta}}_2)$, it holds that

$$\mathrm{E}\left(\mathbb{1}_{T_j \notin [z(\alpha/2);z(1-\alpha/2)]}\right) = \alpha, \quad j = 1, ..., p. \tag{14}$$

That is to say, for repeated samples $\mathcal{S}_1$ and $\mathcal{S}_2 \cup \mathcal{D}$ that are drawn iteratively as decribed before, and for corresponding regression coefficient estimates $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$, we still find significant deviations in $\alpha \cdot 100\%$ of cases under non-informative sampling. However, if $\mathcal{S}_2 \cup \mathcal{D}$ is affected by informative sampling, then $\widehat{\boldsymbol{\beta}}_2$ is biased and $\mathrm{E}(\widehat{\boldsymbol{\beta}}_1) \neq \mathrm{E}(\widehat{\boldsymbol{\beta}}_2)$. In this case, the expectation (15) is larger than α, which indicates systematic deviations stemming from the survey redesign. In other words, we define a survey discontinuity as a situation where

$$\mathrm{E}\left(\mathbb{1}_{T_j \notin [z(\alpha/2);z(1-\alpha/2)]}\right) > \alpha, \quad j = 1, ..., p \tag{15}$$

under the assumption of an underlying constant model over time.

2.4 Correction Methods

We now discuss common methods to account for the bias stemming from missing inclusion probabilities and informative sampling. Recall the diagonal matrix $\boldsymbol{\Pi}_{2\mathcal{D}}$ defined in Section 2.1. This matrix induces a weighting scheme over the observations from the mixed mode survey $\mathcal{S}_2 \cup \mathcal{D}$ for the weighted least squares estimator presented in (10). Due to the lack of information regarding the inclusion probabilities for the sampled individuals $i \in \mathcal{D}$, the term $n_{\mathcal{D}}/N$ was used for weighting. As discussed in the previous section, this choice may not be suitable depending on how $\mathcal{D}$ was collected. In what follows, we show how propensity score estimation and calibration methods can be used to adjust the weighting scheme for situations where the records of the online records are informative.

We start with propensity score estimation. For this, we draw from developments provided by Rosenbaum and Rubin (1983) as well as Lee (2006). However, please note that we partially modify their proposed methods in order to make them applicable to our setting.

The basic idea of applying propensity scores within the statistical framework presented in Section 2.1 is to estimate the unknown inclusion probabilities for all $i \in \mathcal{D}$ by means of a logit model (Nelder and Wedderburn, 1972). That is to say, we assume that there exists a set of real-valued variables $z = \{z_1, ..., z_q\}$ with observed realizations $\mathbf{z}_i \in \mathbb{R}^{1\times q}$ such that $\Pr(i \in \mathcal{D}|\mathbf{z}_i)$ is equal for all individuals. In this case, the logit model describes the log-odds for sample inclusion as a linear function of the z-realizations according to

$$\eta_i(\boldsymbol{\gamma}) := \log \frac{\Pr(i \in \mathcal{D})}{1 - \Pr(i \in \mathcal{D})} = \mathbf{z}_i\boldsymbol{\gamma}. \tag{16}$$

where $\boldsymbol{\gamma} \in \mathbb{R}^{q\times 1}$ is a vector of unknown logit regression coefficients. If (16) holds, then

$$\pi_{\mathcal{D}i} := \Pr(i \in \mathcal{D}) = \frac{\exp(\mathbf{z}_i\boldsymbol{\gamma})}{1 + \exp(\mathbf{z}_i\boldsymbol{\gamma})} = \frac{1}{1 + \exp(-\mathbf{z}_i\boldsymbol{\gamma})} \tag{17}$$

quantifies the unknown inclusion probability for some $i \in \mathcal{D}$, as desired. The remaining step is to estimate the vector of unknown logit regression coefficients. This can be done via maximum likelihood estimation. We minimize the negative loglikelihood under the model

$$\begin{aligned} \widehat{\boldsymbol{\gamma}} &= \underset{\boldsymbol{\gamma}\in\mathbb{R}^q}{\operatorname{argmin}} \; -\log \prod_{i\in\{\mathcal{S}_2\cup\mathcal{D}\}} (\pi_{\mathcal{D}i})^{\mathbb{1}_{(i\in\mathcal{D})}} \\ &= \underset{\boldsymbol{\gamma}\in\mathbb{R}^q}{\operatorname{argmin}} \; -\sum_{i\in\{\mathcal{S}_2\cup\mathcal{D}\}} \left[\mathbb{1}_{(i\in\mathcal{D})}\eta_i(\boldsymbol{\gamma}) - \log\left(1 + \exp(\mathbf{z}_i\boldsymbol{\gamma})\right)\right]. \end{aligned} \tag{18}$$

The solution to (18) can be found numerically, for instance via a Newton-Raphson algorithm. See Train (2009), Chapter 8, for comprehensive insights on the estimation of logit models. Once the logit regression coefficients have been estimated, the inclusion probability for some $i \in \mathcal{D}$ is estimated via model prediction as follows:

$$\widehat{\pi}_{\mathcal{D}i} = \frac{\exp(\mathbf{z}_i\widehat{\boldsymbol{\gamma}})}{1 + \exp(\mathbf{z}_i\widehat{\boldsymbol{\gamma}})} = \frac{1}{1 + \exp(-\mathbf{z}_i\widehat{\boldsymbol{\gamma}})}. \tag{19}$$

The estimated inclusion probability can then be used in order to replace the corresponding main diagonal element of $\boldsymbol{\Pi}_{2\mathcal{D}}$ to adjust the weighting scheme for the weighted least squares estimator presented in (10).

We continue with the calibration approach. For this, we rely on developments provided by Deville and Särndal (1992) and Burgard et al. (2019). Define

$$\boldsymbol{\tau}_{2X} = (\tau_{2X_1}, ..., \tau_{2X_p}) \quad \text{with} \qquad \tau_{2X_j} = \sum_{i\in\mathcal{U}_2} x_{2ij}, \; j = 1, ..., p \tag{20}$$

as covariate population totals and let

$$\widehat{\boldsymbol{\tau}}_{2X} = (\widehat{\tau}_{2X_1}, ..., \widehat{\tau}_{2X_p}), \quad \text{with} \quad \widehat{\tau}_{2X_j} = \sum_{i\in\{\mathcal{S}_2\cup\mathcal{D}\}} w_{2i}x_{2ij}, \; j = 1, ..., p \tag{21}$$

denote the sample estimator of them. Note that (21) is based on the weights w_{2i} that correspond to π_{2i}^{-1} or $N_2/n_{\mathcal{D}}$, depending on whether $i \in \mathcal{S}_2$ or $i \in \mathcal{D}$, respectively. Calibration

in our setting is used to adjust the original weights $w_{21}, ..., w_{2n_2+n_\mathcal{D}}$ such that the population totals (20) are reproduced by the sample estimators (21). The basic idea is that the marginal sample covariate distributions are forced to be consistent with the marginal population covariate distributions. As we are interested in the conditional expectation $\mathrm{E}(Y|X)$, correcting the marginal distribution of X may reduce the bias in regression coefficient estimation when the informativeness of sample inclusion depends on the covariates. For this purpose, we choose a function $D : \mathbb{R} \to \mathbb{R}$ that (implicitely) quantifies the distance between an original weight w_{2i} and an adjusted weight $w_{2i}g_{2i}$, where $g_{2i} \in \mathbb{R}$ is an correction weight. The objective is to minimize the sum over all weight distances while simultaneously ensuring that the estimates in (21) reproduce (20) when the correction weights $g_{21}, ..., g_{2n_2+n_\mathcal{D}}$ are used. That is to say, we solve the constrained optimization problem

$$\min_{\mathbf{g}_2 \in \mathbb{R}^{n_2+n_\mathcal{D}}} \sum_{i \in \{\mathcal{S}_2 \cup \mathcal{D}\}} D(g_{2i}) \quad \text{s.t.} \sum_{i \in \{\mathcal{S}_2 \cup \mathcal{D}\}} w_{2i} g_{2i} x_{2ij} = \tau_{2X_j}, \; j = 1, ..., p, \tag{22}$$

where $\mathbf{g}_2 = (g_{21}, ..., g_{2n_2+n_\mathcal{D}})$. Note that the optimal correction weights, let's say $\widehat{g}_{21}, ..., \widehat{g}_{2n_2+n_\mathcal{D}}$, heavily depend on the choice of D. We use the raking function (Deville and Särndal, 1993)

$$D(g_{2i}) = g_{2i} \log g_{2i} - g_{2i} + 1, \tag{23}$$

as it produces non-negative weights without additional box constraints. The inclusion of box constraints in (22) leads to differentiability problems, as shown by Rupp (2018). Solving the problem then would require quite sophisticated numerical procedures, such as semismooth Newton methods. With the raking function, a standard Newton-Raphson can be used. For further insights on calibration with other distance functions, see Singh and Mohl (1996), as well as Devaud and Tillé (2019). In light of our baseline problem, which is to find a new weighting scheme in order to account for the missing inclusion probabilities, we replace the main diagonal element in $\mathbf{\Pi}_{2\mathcal{D}}$ that corresponds to some sampled individual $i \in \{\mathcal{S}_2 \cup \mathcal{D}\}$ by the term $\widehat{\pi}_{2i} = (w_{2i}\widehat{g}_{2i})^{-1}$.

3 Simulation Study

3.1 Setup

A Monte Carlo simulation study with $R = 1\,000$ iterations indexed by $r = 1, ..., R$ is conducted. We use the synthetic dataset AMELIA on the person-level (Burgard et al., 2017). It contains a realistic artificial population that is generated based on data obtained from the large-scale survey *EU statistics on income and living conditions (EU-SILC)*. See European Commission (2019) for insights on EU-SILC. The AMELIA population consists of 10 012 600 individuals that are hierarchically located in 11 provinces, 40 districts, and 1 592 cities. For the simulation, we draw a random subset of 1 000 000 individuals from the population via simple random sampling. This subset is drawn once prior to the simulation and marks the target population for the subsequent statistical analysis. However, as the AMELIA population is based on a cross-sectional survey, we need to implement an artificial temporal shift for the variables of interest in order to reproduce the statistical framework described in Section 2. The variables of interest are as follows:

- *INC*: personal income / sum of all income variables (Y)
- *PY010*: employee cash or near-cash income (X_1)
- *SOC*: social income (X_2)

We are interested in the statistical relation $Y \sim \beta_1 X_1 + \beta_2 X_2$ within the target population for two time periods. Recall that we assumed the linear model (3) holds for both $t = 1$ and $t = 2$. Therefore, we use the drawn AMELIA subset for both $t = 1$ and $t = 2$ in order to avoid unintended model discontinuities that may result from projecting population from one period into the next. In each iteration of the simulation, we draw samples from the target population for both time periods. For $t = 1$, we draw a 1%-sample of $n_1 = 10\,000$ persons via stratified random sampling. The strata are the 40 districts of AMELIA, while the stratum-specific sample fraction is 1%. This sample represents $\mathcal{S}_1$ in accordance with Section 2. For $t = 2$, we also draw via stratified random sampling in order to obtain $\mathcal{S}_2$. However, the strata are the 11 districts of AMELIA with a proportional stratum-specific sample fraction that varies over simulation scenarios. We let the contribution of $|\mathcal{S}_2| = n_2$ to the total sample size $n_2 + n_{\mathcal{D}} = 10\,000$ vary according to $n_2 \in \{2\,000, 5\,000\}$ in order to study the simulation outcomes under different degrees of augmentation.

Further, we consider four different settings with respect to the informativity of the online records of the web survey. The first is *no informativity*, where $\mathcal{D}$ is drawn via simple random sampling from the target population. The second is *mild informativity*, $\mathcal{D}$ is drawn via simple random sampling from a subgroup of the population. In particular, we only consider individuals with age between 18 and 45. With this selection, some degree of informativity is achieved due to age being positively correlated to income in AMELIA. Next, we have *medium informativity*, where the (unknown) inclusion probabilities are defined as

$$\pi_{2i} = (n_1 - n_2) \cdot \frac{x_{21i}}{\sum_{i \in \mathcal{U}_2} x_{21i}}, \quad \forall\, i \in \mathcal{U}_2 \setminus \mathcal{S}_2. \tag{24}$$

Thus, the inclusion probabilities for the augmenting online data set positively depend on covariate realizations. Therefore, people with larger value for *PY010* have a higher probability of being selected. Finally, we have *strong informativity* by letting the person-specific inclusion probabilities directly depend on the realizations of *INC*:

$$\pi_{2i} = (n_1 - n_2) \cdot \frac{y_{2i}}{\sum_{i \in \mathcal{U}_2} y_{2i}}, \quad \forall\, i \in \mathcal{U}_2 \setminus \mathcal{S}_2. \tag{25}$$

All in all, the simulation scenarios are characterized in Table 1. Under these scenarios, the results of both survey discontinuity evaluation and regression analysis are measured as follows. For the first aspect, we look at the proportion of significant deviations (PSD) in the sense of (13) for a given significance level

$$\mathrm{PSD}(\widehat{\beta}_j, \alpha) = \frac{1}{R} \sum_{r=1}^{R} \mathbb{1}_{(T_j^r \notin [z(\alpha/2), z(1-\alpha/2)])}, \quad \alpha \in \{0.10, 0.05, 0.01\}. \tag{26}$$

Scenario	Sample Sizes	Augmented	Informative	Correction
1	$n_2 = 2\,000$	No	No	No
2	$n_2 = 2\,000$, $n_{\mathcal{D}} = 8\,000$	Yes	No	No
3	$n_2 = 2\,000$, $n_{\mathcal{D}} = 8\,000$	Yes	Mild	No
4	$n_2 = 2\,000$, $n_{\mathcal{D}} = 8\,000$	Yes	Medium	No
5	$n_2 = 2\,000$, $n_{\mathcal{D}} = 8\,000$	Yes	Strong	No
6	$n_2 = 2\,000$, $n_{\mathcal{D}} = 8\,000$	Yes	No	Propensity Score
7	$n_2 = 2\,000$, $n_{\mathcal{D}} = 8\,000$	Yes	Mild	Propensity Score
8	$n_2 = 2\,000$, $n_{\mathcal{D}} = 8\,000$	Yes	Medium	Propensity Score
9	$n_2 = 2\,000$, $n_{\mathcal{D}} = 8\,000$	Yes	Strong	Propensity Score
10	$n_2 = 2\,000$, $n_{\mathcal{D}} = 8\,000$	Yes	No	Calibration
11	$n_2 = 2\,000$, $n_{\mathcal{D}} = 8\,000$	Yes	Mild	Calibration
12	$n_2 = 2\,000$, $n_{\mathcal{D}} = 8\,000$	Yes	Medium	Calibration
13	$n_2 = 2\,000$, $n_{\mathcal{D}} = 8\,000$	Yes	Strong	Calibration
14	$n_2 = 5\,000$	No	No	No
15	$n_2 = 5\,000$, $n_{\mathcal{D}} = 5\,000$	Yes	No	No
16	$n_2 = 5\,000$, $n_{\mathcal{D}} = 5\,000$	Yes	Mild	No
17	$n_2 = 5\,000$, $n_{\mathcal{D}} = 5\,000$	Yes	Medium	No
18	$n_2 = 5\,000$, $n_{\mathcal{D}} = 5\,000$	Yes	Strong	No
19	$n_2 = 5\,000$, $n_{\mathcal{D}} = 5\,000$	Yes	No	Propensity Score
20	$n_2 = 5\,000$, $n_{\mathcal{D}} = 5\,000$	Yes	Mild	Propensity Score
21	$n_2 = 5\,000$, $n_{\mathcal{D}} = 5\,000$	Yes	Medium	Propensity Score
22	$n_2 = 5\,000$, $n_{\mathcal{D}} = 5\,000$	Yes	Strong	Propensity Score
23	$n_2 = 5\,000$, $n_{\mathcal{D}} = 5\,000$	Yes	No	Calibration
24	$n_2 = 5\,000$, $n_{\mathcal{D}} = 5\,000$	Yes	Mild	Calibration
25	$n_2 = 5\,000$, $n_{\mathcal{D}} = 5\,000$	Yes	Medium	Calibration
26	$n_2 = 5\,000$, $n_{\mathcal{D}} = 5\,000$	Yes	Strong	Calibration

Table 1: Characterization of Simulation Scenarios

For regression analysis, we consider bias and variance of model parameter estimation

$$\text{Bias}(\widehat{\beta}_{tj}) = \frac{1}{R}\sum_{r=1}^{R}\widehat{\beta}_{tj}^{r} - \beta_{tj}, \quad \text{Var}(\widehat{\beta}_{tj}) = \frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\beta}_{tj}^{r} - R^{-1}\sum_{s=1}^{R}\widehat{\beta}_{tj}^{s}\right)^2 \tag{27}$$

for $t \in \{1, 2\}$. Here, the true value β_{tj} corresponds to the obtained regression coefficient estimate when considering all individuals from the population of a given period. We further look at the corresponding mean squared error (MSE), which is given by

$$\text{MSE}(\widehat{\beta}_{tj}) = \frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\beta}_{tj}^{r} - \beta_{tj}\right)^2. \tag{28}$$

3.2 Results

We start with survey discontinuity evaluation. They are summarized in Table 2. The table contains the PSD (26) at the significance levels 10%, 5%, and 1%. Note that we restrict the analysis to β_1 in order to avoid confusion resulting from oversized tables. Further, recall equation (15) stating that we expect $\alpha \cdot 100\%$ significant deviations in the absence of informative sampling. Survey discontuities are indicated by surpassing this expectation.

Scenario	**Data**	**Cor.**	**Informative**	**10%**	**5%**	**1%**
	$n_2 = 2\,000$, $n_{\mathcal{D}} = 8\,000$					
1	$\mathcal{S}_2$	No	No	0.119	0.061	0.019
2	$\mathcal{S}_2 \cup \mathcal{D}$	No	No	0.112	0.058	0.013
3	$\mathcal{S}_2 \cup \mathcal{D}$	No	Mild	0.632	0.509	0.281
4	$\mathcal{S}_2 \cup \mathcal{D}$	No	Medium	0.941	0.892	0.758
5	$\mathcal{S}_2 \cup \mathcal{D}$	No	Strong	1.000	1.000	1.000
6	$\mathcal{S}_2 \cup \mathcal{D}$	PS	No	0.137	0.058	0.014
7	$\mathcal{S}_2 \cup \mathcal{D}$	PS	Mild	0.605	0.466	0.233
8	$\mathcal{S}_2 \cup \mathcal{D}$	PS	Medium	0.921	0.857	0.656
9	$\mathcal{S}_2 \cup \mathcal{D}$	PS	Strong	1.000	1.000	1.000
10	$\mathcal{S}_2 \cup \mathcal{D}$	Cal	No	0.129	0.058	0.013
11	$\mathcal{S}_2 \cup \mathcal{D}$	Cal	Mild	0.620	0.482	0.255
12	$\mathcal{S}_2 \cup \mathcal{D}$	Cal	Medium	0.925	0.876	0.724
13	$\mathcal{S}_2 \cup \mathcal{D}$	Cal	Strong	1.000	1.000	1.000
	$n_2 = 5\,000$, $n_{\mathcal{D}} = 5\,000$					
14	$\mathcal{S}_2$	No	No	0.113	0.052	0.015
15	$\mathcal{S}_2 \cup \mathcal{D}$	No	No	0.106	0.051	0.013
16	$\mathcal{S}_2 \cup \mathcal{D}$	No	Mild	0.658	0.541	0.320
17	$\mathcal{S}_2 \cup \mathcal{D}$	No	Medium	0.892	0.836	0.656
18	$\mathcal{S}_2 \cup \mathcal{D}$	No	Strong	1.000	1.000	1.000
19	$\mathcal{S}_2 \cup \mathcal{D}$	PS	No	0.206	0.110	0.040
20	$\mathcal{S}_2 \cup \mathcal{D}$	PS	Mild	0.400	0.248	0.092
21	$\mathcal{S}_2 \cup \mathcal{D}$	PS	Medium	0.785	0.666	0.438
22	$\mathcal{S}_2 \cup \mathcal{D}$	PS	Strong	1.000	1.000	1.000
23	$\mathcal{S}_2 \cup \mathcal{D}$	Cal	No	0.112	0.061	0.017
24	$\mathcal{S}_2 \cup \mathcal{D}$	Cal	Mild	0.457	0.326	0.135
25	$\mathcal{S}_2 \cup \mathcal{D}$	Cal	Medium	0.854	0.778	0.594
26	$\mathcal{S}_2 \cup \mathcal{D}$	Cal	Strong	1.000	1.000	1.000

Table 2: Proportion of Significant Devations per Significance Level

Let us first investigate the overall dependence of PSD and informativity of the augmenting data. In the absence of augmenting data (Scenario 1 and 14), the PSD is between 1.5% and 11.9%, depending on the significance level chosen for the test. By the expectation defined in (15), this is in line with theory. The small deviations from the expected values are due the general Monte Carlo error. In the presence of augmenting data that is non-

informative (Scenario 2, 6, 10, 15, 19, and 23), the PSD ranges from 1.3% to 20.6%, depending on whether a correction method has been used (we address this aspect later). Under informative data augmentation, the PSD figures are significantly larger. Even for mild informativity (Scenario 3, 7, 11, 16, 20, and 24), they range from 9.2% to 63.2%. For medium informativity (Scenario 4, 8, 12, 17, 21, and 25), we have 43.8% to 94.1%. And for strong informativity (Scenario 5, 9, 13, 18, 22, and 26), the PSD is constantly 100%. The findings suggests that as soon as the online records are slightly informative with respect to the model, the outcomes of regression coefficient estimation are significantly different to those obtained from a classical survey sample.

Let us now look at the ability of the correction methods to account for the informativity of the augmenting data. For this, recall that the Scenario 1-5 and 14-18 contain no correction method, Scenario 6-9 and 19-22 implement the propensity score approach, and Scenario 10-13 as well as 23-26 are based on the calibration approach. We see that in the absence of informativity, applying a correction method can actually slightly increase the PSD relative to using no correction. While the PSD with no correction ranges between 1.3% and 11.2%, propensity score estimation obtains 1.4% to 20.6% and the calibration produces 1.3% to 12.9%. This is due to the correction methods implicitly introducing a model to the weighting scheme of the weighted least squares estimator. In the non-informative case without adjustment, the survey weights vary only slightly over the sampled individuals. Applying propensity score estimation increases this variation considerably, as the survey weights now directly depend on the individual covariate realizations. This can lead to large devations in terms of regression coefficient estimation relative to no adjustment. The calibration's impact is not as severe, as in a non-informative sample the population totals are already reproduced in the sense of (22) on expectation. Thus, the necessary weight adjustment for consistency as required in the constrained optimization problem is small.

In the presence of informativity in the augmenting data, a different picture arises. This is visualized in Figure 1. The PSD results with correction are plotted in red, the propensity score results are blue, and the calibration results are marked in green. The horizontal black line displays the expectation under non-informative sampling. We observe that in the presence of mild informativity, the correction methods reduce the PSD by quite a large margin. While no correction has a PSD of 54.1%, the propensity score approach obtains 24.8% and the calibration method yields 32.6%. In the presence of medium informativity, the reduction is still evident, but not as pronounced as before. While no correction has a PSD of 83.6%, propensity score estimation achieves 66.6% and calibration obtains 77.8%. Under strong informativity, we already mentioned that no reduction is evident despite correction. We observe that for mild and medium informativity, the propensity score approach tends to be a slightly better correction method than the calibration method. This could be expected from theory, as propensity score estimation focusses on the baseline problem of our setting, which is the absence of inclusion probabilities for the online records. Calibration focusses on aligning the marginal sample covariate distributions with their counterparts in the population. This only marks an implicit correction method for the problem at hand. By looking back at Table 2, we also see that the correction performance increases with smaller significance levels. Further, it becomes evident that the correction performance is

best in the scenarios where the contribution of the classical survey in terms of sampled individuals is equal to this of the online survey. That is, with a sample size decomposition of $n_2 = 5\,000$, $n_{\mathcal{D}} = 5\,000$, the correction methods can reduce the PSD considerable better than for $n_2 = 2\,000$, $n_{\mathcal{D}} = 8\,000$. However, by looking at the absolute PSD figures, it has to be concluded that overall the correction methods can only slightly mitigate the survey discontinuities resulting from informative sampling. In the majority of estimations, the regression outcomes are significantly different after all.

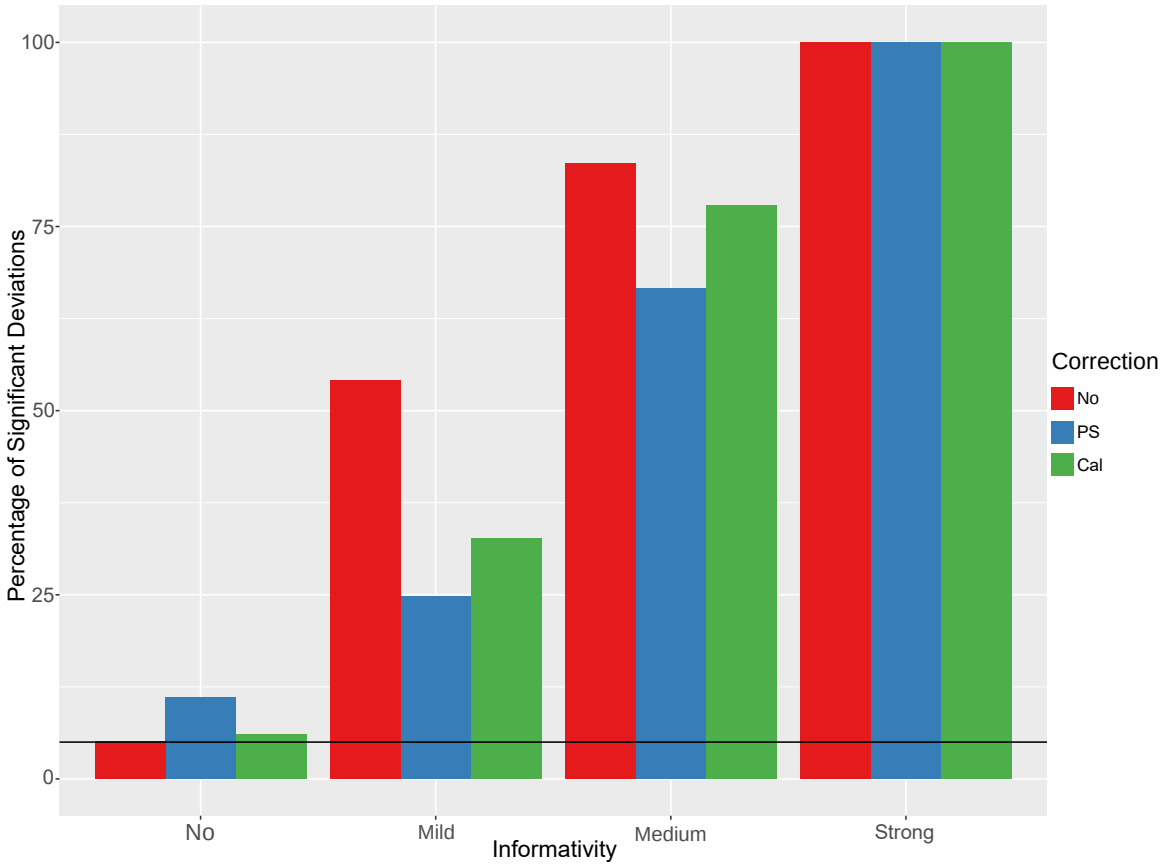

Figure 1: PSD under $\alpha_1 = 0.05$ and $n_2 = 5\,000$, $n_{\mathcal{D}} = 5\,000$

Let us investigate regression coefficient estimation. Again, we focus on the analysis of β_1 in order to avoid oversized tables. The results are summarized in Table 3. Note that we now also include the estimates obtained from the sample $\mathcal{S}_1$ in $t = 1$ as an additional Scenario 0. This is done in order to compare the results within scenarios for the mixed mode survey with the resultsd that would have been achieved under classical survey sampling with decent sample size. First, we observe that overall best estimates are produced based on $\mathcal{S}_1$. The results display the lowest bias and MSE, despite the fact that all considered scenarios (except Scenario 1 and 14) have the same sample size in total. Thus, it can be concluded that the estimates obtained from the mixed mode surveys are less efficient in our setting. The next aspect is that the bias grows with increasing levels of informativity. Given the statistical framework introduced in Section 2.1, this was expected.

Scenario	Data	Cor.	Informative	Bias	Variance	MSE
	$n_2 = 2\,000$, $n_{\mathcal{D}} = 8\,000$					
0	$\mathcal{S}_1$	No	No	0.00008	0.00010	0.00010
1	$\mathcal{S}_2$	No	No	0.00063	0.00047	0.00047
2	$\mathcal{S}_2 \cup \mathcal{D}$	No	No	-0.00018	0.00010	0.00010
3	$\mathcal{S}_2 \cup \mathcal{D}$	No	Mild	-0.02651	0.00011	0.00081
4	$\mathcal{S}_2 \cup \mathcal{D}$	No	Medium	-0.03239	0.00010	0.00115
5	$\mathcal{S}_2 \cup \mathcal{D}$	No	Strong	-2.10331	0.03882	4.46274
6	$\mathcal{S}_2 \cup \mathcal{D}$	PS	No	-0.00021	0.00011	0.00011
7	$\mathcal{S}_2 \cup \mathcal{D}$	PS	Mild	-0.02557	0.00010	0.00076
8	$\mathcal{S}_2 \cup \mathcal{D}$	PS	Medium	-0.02887	0.00007	0.00090
9	$\mathcal{S}_2 \cup \mathcal{D}$	PS	Strong	-0.33197	0.00044	0.11064
10	$\mathcal{S}_2 \cup \mathcal{D}$	Cal	No	-0.00018	0.00010	0.00010
11	$\mathcal{S}_2 \cup \mathcal{D}$	Cal	Mild	-0.02609	0.00011	0.00079
12	$\mathcal{S}_2 \cup \mathcal{D}$	Cal	Medium	-0.03186	0.00013	0.00114
13	$\mathcal{S}_2 \cup \mathcal{D}$	Cal	Strong	-1.92951	0.03623	3.75924
	$n_2 = 5\,000$, $n_{\mathcal{D}} = 5\,000$					
14	$\mathcal{S}_2$	No	No	0.00066	0.00019	0.00019
15	$\mathcal{S}_2 \cup \mathcal{D}$	No	No	0.00013	0.00010	0.00010
16	$\mathcal{S}_2 \cup \mathcal{D}$	No	Mild	-0.02515	0.00021	0.00084
17	$\mathcal{S}_2 \cup \mathcal{D}$	No	Medium	-0.03059	0.00013	0.00107
18	$\mathcal{S}_2 \cup \mathcal{D}$	No	Strong	-1.20945	0.03986	1.50263
19	$\mathcal{S}_2 \cup \mathcal{D}$	PS	No	-0.00033	0.00020	0.00020
20	$\mathcal{S}_2 \cup \mathcal{D}$	PS	Mild	-0.01671	0.00012	0.00040
21	$\mathcal{S}_2 \cup \mathcal{D}$	PS	Medium	-0.02359	0.00010	0.00065
22	$\mathcal{S}_2 \cup \mathcal{D}$	PS	Strong	-0.27848	0.00070	0.07825
23	$\mathcal{S}_2 \cup \mathcal{D}$	Cal	No	0.00016	0.00010	0.00010
24	$\mathcal{S}_2 \cup \mathcal{D}$	Cal	Mild	-0.01800	0.00013	0.00045
25	$\mathcal{S}_2 \cup \mathcal{D}$	Cal	Medium	-0.02463	0.00012	0.00073
26	$\mathcal{S}_2 \cup \mathcal{D}$	Cal	Strong	-1.15623	0.03703	1.37390

Table 3: Results of Model Parameter Estimation

However, the correction methods are capable of reducing the bias to a notable extent, especially for the sample size decomposition of $n_2 = 5\,000$, $n_{\mathcal{D}} = 5\,000$. Let

$$\frac{\text{Bias}(\widehat{\beta}_1^{no}) - \text{Bias}(\widehat{\beta}_1^{ad})}{\text{Bias}(\widehat{\beta}_1^{no})} \cdot 100\%$$

denote the relative reduction in percent achieved by correction. Under mild informativity, the propensity score approach reduces the bias between 4% and 51%. The calibration method yields 2% to 43%. For medium informativity, we have 11% to 23% and 2% to 19%, respectively. The largest bias reduction is achieved under strong informativity. Here, the propensity score method obtains a reduction between 77% and 84%. The calibration

approach achieves 4% to 8%. Thus, even though the hypothesis tests based on (12) and (13) display a very high share of significant deviations under informativity despite correction, the overall $\boldsymbol{\beta}$-inference is considerably improved by reweighting.

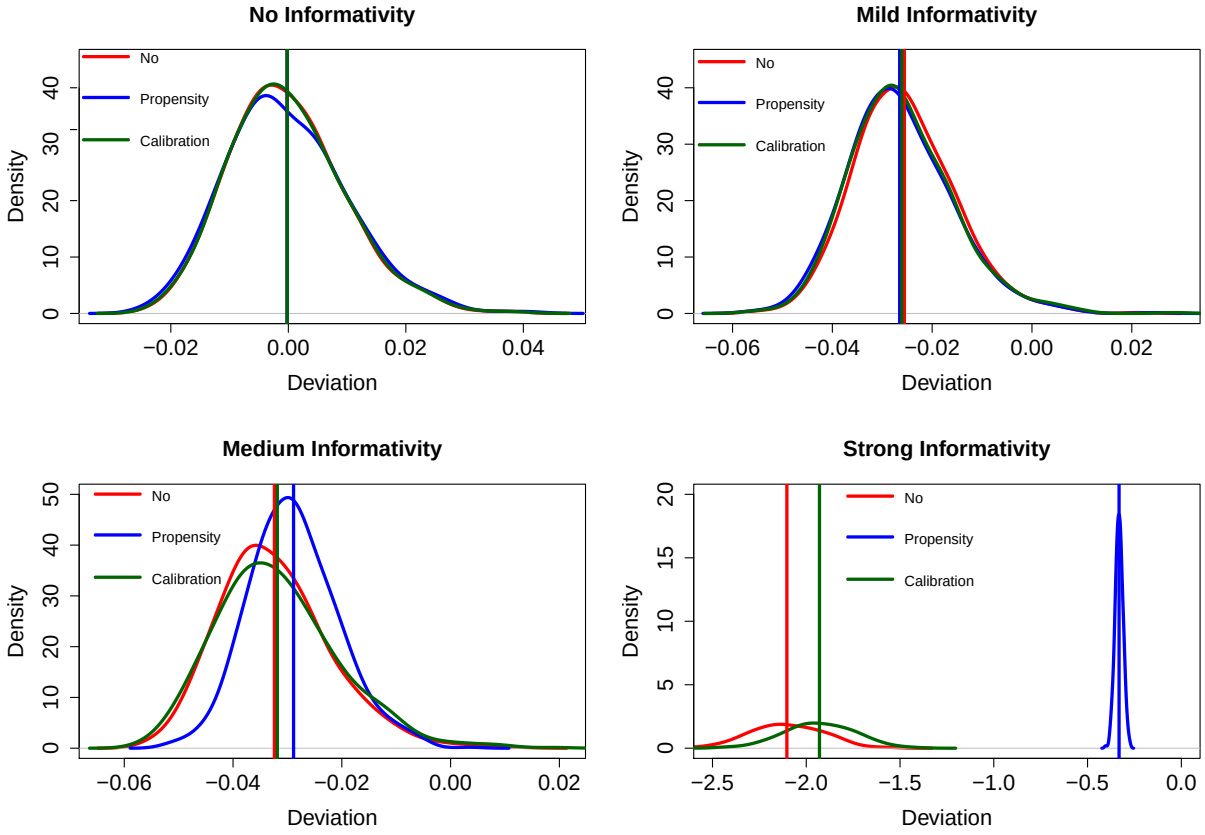

Figure 2: Deviation of Model Parameter Estimation, $n_2 = 2\,000$, $n_{\mathcal{D}} = 8\,000$

This is further visualized in Figure 2. It displays the densities of the devation $\widehat{\beta}_1 - \beta_1$ over all simulation iterations and degrees of informativity. Again, the results without correction are plotted in red, those obtained by propensity score estimation are blue, and the results of calibration are marked in green. The graph supports the bias reduction tendencies of both correction methods mentioned before. It further supports the finding that propensity score estimation is the better correction method for our setting, which was already evident for the survey discontinuities. Another interest observation is that the propensity score is capable of reducing the estimation variance in the presence of informativity. Under medium and strong informativity, we see that the blue densities are not only closer located to zero, their overall masses are also more concentrated around their respective centers of gravity.

4 Conclusion and Outlook

We studied survey discontinuities in settings where a classical survey sample is substituted by a mixed mode survey that relies on web-augmentation in terms of online records. On the

example of linear regression, we investigated how inference regarding the regression coefficients is affected by a corresponding change in data collection. For this purpose, a suitable hypothesis test was presented that assesses whether the outcomes of regression analysis in the two surveys are significantly different. A special emphasis was placed on situations where the records of the online survey are informative with respect to the regression model. We further discussed propensity score estimation and calibration as potential methods for correcting the bias resulting from a potential informativity of the augmenting data. An extended Monte Carlo simulation study was conducted in order to assess the effects of the mentioned survey redesign under different degrees of informativity. We found that even mild informativity of the augmenting data leads to survey discontinuities in the majority of cases. It further impairs the results of regression coefficient estimation considerably. The presented correction methods are capable of reducing the negative effects of informative online records to some extent. However, the overall quality of estimates obtained from the classical survey cannot be achieved by any means.

The presented paper makes a case on treating online data sources in the context of survey analysis carefully. Web-augmented mixed mode surveys undoubtedly have great advantages. They are overall resource-efficient and – depending on the application – even allow to empirically investigate areas of life that are typically hard to monitor via classical surveys. Therefore, they indeed mark a valuable addition to socioeconomic and political research in future studies. However, since model-based inference has emerged as the primary approach to quantitative analysis in these fields, researchers have to carefully evaluate whether the data bases are informative with respect to their models. Currently, much research effort is put in finding suitable correction methods for bias stemming from informativity. Yet, as our simulation study suggest, the effectiveness of such approaches very much depends on the degree of informativity and on the availability of suitable auxiliary data for correction.

Acknowledgements

This research was conducted within the research project *MAKSWELL - Making sustainable development and well-being frameworks work for policy* in the course of the *Horizon 2020* programme funded by the European Union. We kindly thank for the financial support.

References

Burgard, J. P., J.-P. Kolb, H. Merkle, and R. Münnich (2017). Synthetic data for open and reproducible methodological research in social sciences and official statistics. *AStA Wirtschafts- und Sozialstatistisches Archiv 11*, 233–244.

Burgard, J. P., R. Münnich, and M. Rupp (2019). A generalized calibration approach ensur-

ing coherent estimates with small area constraints. *Research Papers in Economics 10/19*. Trier University.

Cassel, C. M., C. E. Särndal, and J. H. Wretman (1977). *Foundations of inference in survey sampling*. New York: Wiley & Sons.

Clogg, C. C., E. Petkova, and A. Haritou (1995). Statistical methods for comparing regression coefficients between models. *American Journal of Sociology 100*(5), 1261–1293.

Devaud, D. and Y. Tillé (2019). Deville and särndal's calibration: Revisiting a 25-years-old successful optimization problem. *TEST 28*, 1033–1065.

Deville, J.-C. and C.-E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association 87*(418), 376–382.

Deville, J.-C. and C.-E. Särndal (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association 88*(423), 1013–1020.

European Commission (2019). EU statistics on income and living conditions (EU-SILC) methodology. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=EU_statistics_on_income_and_living_conditions_(EU-SILC)_methodology.

Fuller, W. A. (2009). *Sampling statistics*. Hoboken, New Jersey: Wiley & Sons.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics 22*(2), 329–349.

Nelder, J. and R. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General) 135*(3), 370–384.

Pasternoster, R., R. Brame, P. Mazerolle, and A. Piquero (1998). Using the correct statistical test for the equality of regression coefficients. *Criminology 36*(4), 859–866.

Pfeffermann, D. and M. Sverchkov (2009). Inference under informative sampling. In C. R. Rao (Ed.), *Handbook of Statistics: Sample Surveys: Inference and Analysis*, Volume 28, pp. 455–487. Elsevier.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rupp, M. (2018). Optimization for multivariate and multi-domain methods in survey statistics. PhD Thesis, Trier University.

Singh, A. and C. Mohl (1996). Understanding calibration estimators in survey sampling. *Survey Methodology 22*, 107–115.

Smith, P. A., N. Tzavidis, T. Schmid, N. Rojas, and J. van den Brakel (2017). Identifying potential discontinuities in the new national survey for wales. Technical report.

Särndal, C. E., B. Swensson, and J. Wretman (1992). *Model assisted survey sampling*. New York: Springer.

Train, K. E. (2009). *Discrete choice methods with simulation* (2 ed.). New York: Cambridge University Press.

van den Brakel, J. and J. Roels (2010). Intervention analysis with state-space models to estimate discontinuities due to a survey redesign. *The Annals of Applied Statistics 4*(2), 1105–1138.

van den Brakel, J., P. A. Smith, and S. Compton (2008). Quality procedures for survey transitions - experiments, time series and discontinuities. *Survey Research Methods 2*(3), 123–141.

van den Brakel, J., M. Zhang, and S.-M. Tam (2020). Measuring discontinuities in time series obtained with repeated sample surveys. *International Statistical Review*. Online-first version.

Zagheni, E. and I. Weber (2015). Demographic research with non-representative internet data. *International Journal of Manpower 36*(1), 13–25.