TI 2024-066/III
Tinbergen Institute Discussion Paper

Mitigating Estimation Risk: a Data-Driven Fusion of Experimental and Observational Data

Francisco Blasques[1]
Paolo Gorgi[2]
Siem Jan Koopman[3]
Noah Stegehuis[4]

1 Vrije Universiteit Amsterdam, Tinbergen Institute
2 Vrije Universiteit Amsterdam, Tinbergen Institute
3 Vrije Universiteit Amsterdam, Tinbergen Institute
4 Vrije Universiteit Amsterdam, Tinbergen Institute

Mitigating Estimation Risk: A Data-Driven Fusion of Experimental and Observational Data*

F. Blasques[1, 2], *P. Gorgi*[1, 2], *S.J. Koopman*[1, 2], and *N. Stegehuis*[1, 2]

[1]Department of Econometrics & Data Science, Vrije Universiteit Amsterdam, The Netherlands
[2]Tinbergen Institute, The Netherlands

Abstract

The identification of causal effects of marketing campaigns (advertisements, discounts, promotions, loyalty programs) require the collection of experimental data. Such data sets frequently suffer from limited sample sizes due to constraints (time, budget) which can result in imprecise estimators and inconclusive outcomes. At the same time, companies passively accumulate observational data which oftentimes cannot be used to measure causal effects of marketing campaigns due to endogeneity issues. In this paper we show how estimation uncertainty of causal effects can be reduced by combining the two data sources by employing a self-regulatory weighting scheme that adapts to the underlying bias and variance. We also introduce an instrument-free exogeneity test designed to assess whether the observational data is significantly endogenous and experimentation is necessary. To demonstrate the effectiveness of our approach, we implement the combined estimator for a real-life data set in which returning customers were awarded with a discount. We demonstrate how the indecisive result of the experimental data alone can be improved by our weighted estimator, and arrive to the conclusion that the loyalty discount has a notably negative effect on net sales.

Keywords: endogeneity, data fusion, experimental data, observational data

*Corresponding author: Noah Stegehuis (n.stegehuis@vu.nl). Francisco Blasques and Noah Stegehuis acknowledge the financial support of the Dutch Science Foundation (NWO) under grant Vidi.195.099.

1 Introduction

Recent digitalisation has shifted decision-making in businesses from expert-driven judgement to data-driven methods. Additionally, increasing opportunities in data collection have allowed firms in digital marketing to passively accumulate fine-grained data of customer decisions such as online purchases and ad click-through rates. The aim of much quantitative marketing research is to measure the causal effect of product and marketing design choices like discounts, prices, advertising campaigns, promotions, product recommendations, new product features, etc. The use of observational data sets is problematic for measuring causal effects, as such data may suffer from endogeneity issues that originate from omitted variables, simultaneity and measurement errors in the variables. All such issues may lead to a bias in the estimator of the causal impact (Lewis et al. 2011). Literature reviews on endogeneity issues, their sources, implications and remedies in marketing research are presented in Rutz and Watson (2019), Papies et al. (2017), Ebbes et al. (2022).

A well-designed experimental data set is much less prone to endogeneity issues (Kohavi et al. 2009, 2020). Major companies such as Facebook (Gordon et al. 2019), Microsoft (Li et al. 2015) and Amazon (Cui et al. 2019) have executed experiments to measure causal effects of design choices on consumer decisions. The main challenge with experimental data is that it is often limited in scale, as the design and execution of experiments are commonly considered expensive and operationally inefficient (Gluck 2011). Furthermore, marketing budgets are limited in practice and obtaining large enough data sets would require a lot of commitment and patience (Campbell et al. 2022). Finally, firms may be confronted with ethical questions when they offer different prices or promotions to similar customer groups for an extended periods of time while, at the same time, they may run the risk of harming customer satisfaction (Nunan and Di Domenico 2022). These considerations are especially relevant for small-to-medium businesses. Hence, for such companies it is challenging to acquire experimental data sets with large enough sample sizes. As a result, there is no convincing evidence for the (in)effectiveness of the marketing campaign since these small data sets typically do not have enough power to measure an effect.

While observational and small experimental data sets individually have their drawbacks, combining them can potentially be highly beneficial. In this paper, we propose a method of combining the two data sources, using a weighted average of the separate estimates of the causal impact. The weight is self-regulatory as it adapts to the bias-variance trade-off present in the data. For instance, when the observational data produces a largely biased estimate due to endogeneity, the experimental estimate is then relatively more reliable and the weight will favor the latter estimate. Conversely, whenever the experimental sample size is small and the corresponding estimate is relatively less efficient, the weight will now favor the experimental estimate less. After introducing this new estimation method and analysing its asymptotic properties, we propose a test for exogeneity to formally assess the presence of endogeneity in the observational data. The proposed test is similar to the Durbin-Wu-Hausman test (see Durbin (1954), Wu (1973) and Hausman (1978)). However, in contrast, our exogeneity test does not rely on instrumental variables, but exploits the exogeneity of the experimental data. This test can prove to be useful in practice, as a rejection implies that experimentation might indeed be

necessary to obtain an unbiased result.

To illustrate our method we consider an experimental-observational data set from a Dutch phone repair company (ThePhoneLab) that is interested in the effect of a loyalty discount scheme based on an annual allowance for returning customers. This eligibility procedure induces customers to be self-selected into treatment. Such a scheme is likely giving rise to endogeneity due to omitted variables. Therefore, to be able to measure the true impact of the discount, an experiment has been conducted for a short time period in which customers were randomly selected to be eligible for the discount. The resulting experimental estimate has suffered from a large variance and has led to an inconclusive result about the effect of the discount on total net sales. When applying our proposed estimation and testing procedures to this case, we conclude that there is in fact a significant negative effect. More specifically, while the company was not able to measure the effect of the discount using experimental data only, our method has established that the discount was in fact harmful. The discount scheme did not provide sufficient incentives to customers to buy more than the cost of the discount. This case has also demonstrated that our method is straightforward and simple, and it can be implemented without relying on complicated statistical methods.

The remainder of the paper is structured as follows. In Section 2 we review the relevant literature. In Section 3 we introduce the estimator, discuss its asymptotic properties and propose the exogeneity test. In Section 4 we show the MSE reduction by means of a simulation study, as well as a power and size analysis of the exogeneity test. Section 5 demonstrates the use of the estimator in the loyalty discount application, followed by the conclusion in Section 6.

2 Literature Review and Research Contribution

The advantages of combining observational and experimental data have already been widely discussed and explored in the recent literature within various frameworks. For example, Cooper and Yoo (2013), Fernandez Loria and Provost (2020) and Gasse et al. (2021) integrate the two data sources into machine learning algorithms to establish causal relationships. Athey et al. (2020) and Rosenman et al. (2023) use experimental and observational data to identify (long-term) treatment effects. Rosenman et al. (2023) also apply a shrinkage weight to combine observational and experimental estimators, which in their case contain the treatment effects of each strata. Our context is a more general regression set up and we combine the estimators of both data sources directly by means of a data-driven weighting scheme. This current research project is closely related to Gui (2024) who considers a general regression framework. Although this paper focuses on a generalised method of moments (GMM) approach, with an extra moment condition associated with an imperfect (potentially endogenous) instrument for the observational data, it also discusses reducing the variance by weighing the (bias-corrected) estimators according to a scheme that minimises a mean squared error (MSE) criterion. In our research project, we adopt a similar estimation strategy but combine multivariate estimates without bias-correction. We show that the resulting weights adjust automatically to the existing variance and bias present in the data. Furthermore, we investigate the asymptotic properties and provide a simulation study to

analyse small-sample behavior.

We draw on the statistical literature (Green and Strawderman 1991, Judge and Mittelhammer 2004, Mittelhammer and Judge 2005) for our weighting scheme. Although these papers proposed the weighing of estimates to combine any type of estimate, the weighing method has also been used for combining experimental and observational estimates specifically Rosenman et al. (2023). However, the Stein-like (SL) weight considered in these works typically jumps to 0 when the two estimates have values close to each other, implying in this setting that the experimental data is excluded altogether. This case can even occur if there is a bias in the observational estimator, as the large variance of the experimental estimator can still initiate the zero weight. In such cases it would be quite harmful to rely on the observational data alone. Therefore, we consider an alternative weight that is less sensitive and never fully eliminates an estimator. This weight is more conservative than the SL weight, and hence does not favor the observational data on the outset, and it shows an improved MSE performance in cases where endogeneity is considerably present.

The relevance of our proposed exogeneity test is shown in the work of Gordon et al. (2019) where it is investigated for various relevant data sets whether the measured causal effect of digital advertising is different when based on experimental or observational data. In their research it is concluded that for the majority of cases there is a significant bias in the observational data. Our proposed simple test provides a quick method for detecting endogeneity by considering the scaled difference between the observational and experimental estimates.

3 Modeling Framework, Estimation and Testing

We consider a regression modeling framework and assume that y, the data variable of interest, is generated by the linear regression model[1] as given by

$$y_i = \alpha + \boldsymbol{\beta}_0' \boldsymbol{x}_i + \varepsilon_i, \tag{1}$$

where y_i is the i-th realisation for variable y, α is the unknown intercept or constant, $\boldsymbol{\beta}_0$ is the unknown $k \times 1$ parameter vector of interest, $\boldsymbol{x}_i$ is a vector of k regressors, ε_i is the disturbance term that is identically and independently distributed with mean zero and variance σ^2. We distinguish between the experimental and observational data sets by indexing the data and corresponding estimators by their respective sample sizes. The experimental data set is based on sample size N and is denoted by y_i^N for $i = 1, \ldots, N$, that is generated by (1) with exogeneous regressors $\boldsymbol{X}_N = (\boldsymbol{x}_1^N, \ldots, \boldsymbol{x}_N^N)'$ so that $\mathbb{E}[\varepsilon_i^N | \boldsymbol{x}_i^N] = 0$. Based on this data set we obtain the experimental estimator, denoted by $\hat{\boldsymbol{\beta}}_N$, by means of estimating the model in (1) with ordinary least squares (OLS). We define the observational data set with sample size T analogously by y_i^T for $i = 1, \ldots, T$, and is generated by the same model (1), but

[1] The data generating process can be more complex. For example, the model can include nonlinear functions of the regressors. It requires the use of appropriate nonlinear (least squares) variance estimators rather than ordinary least squares estimators in our current discussion. Our method only requires the estimators themselves and their variance matrices.

with endogeneous regressors $\boldsymbol{X}_T = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_T^T)'$ in the sense that $\mathbb{E}[\varepsilon_i^T | \boldsymbol{x}_i^T] \neq 0$. The observational estimator is denoted as $\hat{\boldsymbol{\beta}}_T$. We are particularly interested in the case where $T \gg N$.

3.1 The Causal Observational Shrinkage Estimator

The *Causal-Observational Shrinkage Estimator* (COSE) is defined as the linear combination of the experimental and observational estimators with some weight $\lambda \in [0, 1]$, that is

$$\tilde{\boldsymbol{\beta}}_{NT}(\lambda) = \lambda \hat{\boldsymbol{\beta}}_N + (1 - \lambda)\hat{\boldsymbol{\beta}}_T. \tag{2}$$

We aim to find an estimator that minimises the quadratic risk, also referred to as mean squared error (MSE). The MSE provides a way to take into account both bias and variance and is defined for some estimator $\hat{\boldsymbol{\beta}}$ as

$$\text{MSE}(\hat{\boldsymbol{\beta}}) = \mathbb{E}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 = \text{bias}^2 + \text{variance}.$$

To allow for a larger variety of estimation techniques, we do not make distributional assumptions on the errors, and only require that the estimators are asymptotically normal and have some estimable variance matrix, as formalised in Assumptions 1 and 2. We denote the limiting bias of the observational estimator as $\boldsymbol{\gamma}_0$, which is possibly equal to zero. In the case of OLS estimators and independently normally distributed errors, $\boldsymbol{\Sigma}_N$ in Assumption 1 is simply $\boldsymbol{\Sigma}_N = \sigma^2 \left(\boldsymbol{X}'_N \boldsymbol{X}_N\right)^{-1}$, which is estimated by $\hat{\boldsymbol{\Sigma}}_N = s^2 \left(\boldsymbol{X}'_N \boldsymbol{X}_N\right)^{-1}$ with $s^2 = \boldsymbol{e}'_N \boldsymbol{e}_N \,/\, N - k$ and where $\boldsymbol{e}_N$ is the vector of residuals of the experimental regression.

Assumption 1. *The k-dimensional experimental estimator is unbiased and distributed with some variance matrix $\hat{\boldsymbol{\beta}}_{\boldsymbol{N}} \sim (\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_N)$, and is asymptotically normally distributed with $\sqrt{N}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_0)$ where $\boldsymbol{\Sigma}_0 = \operatorname*{plim}_{N\to\infty} N\hat{\boldsymbol{\Sigma}}_N$.*

Assumption 2. *The (potentially) biased k-dimensional observational estimator is distributed with some variance matrix $\hat{\boldsymbol{\beta}}_T \sim (\boldsymbol{\beta}_{\boldsymbol{0}} + \boldsymbol{\gamma}_T, \boldsymbol{\Phi}_T)$ where the bias $\boldsymbol{\gamma}_T \to \boldsymbol{\gamma}_0$ as $T \to \infty$ for some $\boldsymbol{\gamma}_0 \in \mathbb{R}$. We also assume it is asymptotically normally distributed $\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0 - \boldsymbol{\gamma}_0) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Phi}_0)$, where $\boldsymbol{\Phi}_0 = \operatorname*{plim}_{T\to\infty} T\hat{\boldsymbol{\Phi}}_T$.*

Judge and Mittelhammer (2004) show that by minimising the MSE of $\tilde{\boldsymbol{\beta}}_{NT}(\lambda)$ with respect to the weight λ, we get an optimal weight that is an expression of the bias and variances of the estimators. In general terms, we have

$$\lambda^* = 1 - \frac{\text{tr}(\boldsymbol{\Sigma}_N)}{\text{tr}(\boldsymbol{\Sigma}_N) + \text{tr}(\boldsymbol{\Phi}_T) + \boldsymbol{\gamma}'_T \boldsymbol{\gamma}_T}. \tag{3}$$

This weight is infeasible, as it contains the true unknown expressions of $\boldsymbol{\Sigma}_N$, $\boldsymbol{\gamma}_T$ and $\boldsymbol{\Phi}_T$. They show that a feasible estimator of this weight is given by

$$\hat{\lambda}^{SL} = 1 - \frac{\text{tr}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{N}})}{\left\|\hat{\boldsymbol{\beta}}_N - \hat{\boldsymbol{\beta}}_T\right\|^2}, \tag{4}$$

since the denominator is in expectation equal to the denominator of the optimal weight in (3), i.e. $\mathbb{E}\|\hat{\boldsymbol{\beta}}_N - \hat{\boldsymbol{\beta}}_T\|^2 = \text{tr}(\boldsymbol{\Sigma}_N) + \text{tr}(\boldsymbol{\Phi}_T) + \boldsymbol{\gamma}_T'\boldsymbol{\gamma}_T$. It turns out that with normally distributed error terms and under the optimal weight λ^*, the MSE of the weighted estimator is never worse than the MSE of the base estimator whenever the number of regressors $k \geq 5$ (Judge and Mittelhammer 2004).

After plugging the estimators into the estimated weight $\hat{\lambda}^{SL}$, the dominance of the finite-sample MSE of the COSE cannot be guaranteed. However, an approximation provides an insight in how this dominance is plausible. The finite sample MSE can be expanded up to first order, and the resulting approximation has a strictly lower MSE. More specifically, we have

$$MSE(\tilde{\boldsymbol{\beta}}_{NT}(\hat{\lambda}^{SL})) \approx \text{tr}\left(\boldsymbol{\Sigma}_N\right) \times \left(1 - \frac{\text{tr}\left(\boldsymbol{\Sigma}_N\right)}{\mathbb{E}\|\hat{\boldsymbol{\beta}}_N - \hat{\boldsymbol{\beta}}_T\|^2}\right). \tag{5}$$

It is clear that the last term on the right-hand side of equation (5) (which is strictly smaller than 1) brings about a smaller risk than the experimental estimator with a constant risk of $\text{MSE}(\hat{\boldsymbol{\beta}}_N) = \text{tr}\left(\boldsymbol{\Sigma}_N\right)$. However, in practice the behavior of the $\hat{\lambda}^{SL}$ weight is erratic whenever the point estimates are close together, such that $\|\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_N\|^2$ is near 0. The near zero difference in the denominator causes the fraction to become excessively large, leading to a highly negative value of $\hat{\lambda}^{SL}$. Since it does not naturally fall in the [0,1] interval, a common solution is to force any negative number to 0[2]. So instead of $\hat{\lambda}^{SL}$ we consider $\min\{0, \hat{\lambda}^{SL}\}$, in line with Judge and Mittelhammer (2004) and Rosenman et al. (2023).

The zero weight due to similar point estimates could occur when *i)* observational data is (nearly) exogeneous, or *ii)* due to the large variance of the experimental estimator which could incidentally produce a value close to the observational estimator. In setting *i)* it is acceptable to put all weight on the observational data since it is unbiased. However, in setting *ii)* it would be harmful to rely solely on observational data when there is in fact a large bias present. In empirical research, we are faced with a one-off estimation problem, we cannot know which of the two cases is the underlying force that results in the zero weight. Furthermore, the zero SL weight that occurs when the observational data is exogenous is actually undesirable. In the extreme case where the observational data is drawn from the same distribution as the experimental data, the variances are identical. This leads to an optimal weight of $\lambda^* = 0.5$ since $\boldsymbol{\gamma}_T = 0$ and $\boldsymbol{\Sigma}_N = \boldsymbol{\Phi}_T$. It also makes intuitive sense that each is weighed equally, as both estimators are reliable, and taking both into account yields a lower overall variance.

[2]This estimator is actually closely related to the James-Stein (JS) estimator (Stein 1956, James and Stein 1961) that shrinks an estimator towards zero, which is proven to reduce the MSE. It is a well known that the positive-part JS estimator that rules out negative weights, is an improvement of the regular JS estimator (Anderson 1984).

It would be appropriate that as more data becomes available, the estimated weight also converges to the optimal 0.5. But, even in the limit, $\hat{\lambda}^{SL}$ will jump to 0 due to the zero denominator. Also in more moderate cases, where the design matrices are similar due to $\boldsymbol{\gamma}_0 = 0$ and $N = T$, the optimal weight will be near 0.5 while the SL weight often excludes the experimental data altogether.

In an attempt to resolve these issues, we consider the equally valid, more conservative estimator

$$\hat{\lambda}^{C} = 1 - \frac{\operatorname{tr}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{N}})}{\operatorname{tr}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{N}}) + \operatorname{tr}(\hat{\boldsymbol{\Phi}}_{\boldsymbol{T}}) + \|\hat{\boldsymbol{\beta}}_N - \hat{\boldsymbol{\beta}}_T\|^2} \tag{6}$$

which also appears in Mittelhammer and Judge (2005). However, we highlight the value of this weight in the context of combining experimental and observational data. Most importantly, it is less sensitive to nearly equal point estimates $\hat{\boldsymbol{\beta}}_N$ and $\hat{\boldsymbol{\beta}}_T$, meaning it is less likely to exclude the valuable experimental data. Furthermore, it is naturally bounded within the $[0, 1]$ interval and in the case of equal datasets yields the value of 0.5 rather than 0.

3.2 Asymptotic Properties

We investigate the limiting properties of the COSE in equation (6) under the nonrestrictive Assumptions 1 and 2 that both estimators are asymptotically normally distributed and are consistent for the true parameter $\boldsymbol{\beta}_0$ and pseudo-true parameter $(\boldsymbol{\beta}_0 + \boldsymbol{\gamma}_0)$ respectively.

We show in Proposition 1 that the causal-observational shrinkage estimator always converges to the true unknown parameter $\boldsymbol{\beta}_0$, as the experimental sample size increases, for both the SL and the conservative weight. This is a direct result of the MSE minimising weights, that automatically adjust to 1 when the observational data is biased and as enough experimental data comes available ($N \to \infty$), meaning that the COSE converges to the reliable and efficient estimator $\hat{\boldsymbol{\beta}}_N$. The consistency result even holds when there is no bias in the observational data ($\boldsymbol{\gamma}_0 = 0$) and is valid for any growth rate of T relative to N, whether it is fixed ($\alpha = 0$), grows faster ($\alpha > 1$), equally fast ($\alpha = 1$) or even slower ($\alpha < 1$) than N. Proofs are given in the Appendix.

Proposition 1 (Consistency)**.** *Let T grow relative to N by the relation $T = cN^{\alpha}$ for some constant c and $\alpha \geq 0$. Let the limiting bias $\boldsymbol{\gamma}_0$ be any value including zero. Then, under Assumptions 1 and 2 the causal-observational shrinkage estimator is consistent for $\boldsymbol{\beta}_0$, i.e.*

$$\tilde{\boldsymbol{\beta}}_{NT}\left(\hat{\lambda}\right) \xrightarrow{p} \boldsymbol{\beta}_0 \qquad \text{as } N \to \infty \tag{7}$$

for both $\hat{\lambda}^{SL}, \hat{\lambda}^{C}$.

Remark 1. *Note that Proposition 1 includes as a special case our large-observational-small-experimental scenario of interest, where N is not only a small fraction of T (large c), but possibly also diverges to infinity at a rate arbitrarily slower than T ($\alpha > 1$). The proposition even includes the case of a large experiment where T is fixed ($\alpha = 0$) and only N grows to infinity.*

Denote the limiting distribution of the experimental estimator by $\mathcal{D} := \lim_{N\to\infty} \sqrt{N}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0) \overset{d}{=} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_0)$. Proposition 2 states that with a non-zero limiting bias COSE has the same limiting distribution $\mathcal{D}$, regardless of how fast the observational and experimental datasets increase relative to each other (T fixed, same rate, T increasing faster or slower) as a result of a vanishing weight. The weight vanishes, as the experimental estimator with a large sample size is highly efficient, and including the biased observational data would only increase the MSE. This shows the self-regulatory behaviour of the weights, that provide the optimal minimum MSE in the limit. Although one would expect that scaling the biased COSE by $\sqrt{N}$ will make the term explode, this does not happen as the squared norm in the denominator of the weight will grow faster hence reducing the weight to zero more rapidly than the scaled estimator can blow up.

Proposition 2 (Asymptotic Distribution non-zero bias)**.** *Let the limiting bias $\boldsymbol{\gamma}_0 \neq 0$, for some $\boldsymbol{\gamma}_0 \in \mathbb{R}$ and $T = cN^\alpha$, for some constant c and $\alpha \geq 0$. Then, under Assumptions 1 and 2*

$$\sqrt{N}\left(\tilde{\boldsymbol{\beta}}_{NT}\left(\hat{\lambda}\right) - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathcal{D} \qquad \text{as } N \to \infty \tag{8}$$

for both $\hat{\lambda}^{SL}, \hat{\lambda}^{C}$.

Although in most practical cases observational data will be endogeneous and yield a biased estimator as is considered in Proposition 2, it is important to investigate the asymptotic properties when there is no endogeneity. Proposition 3 shows that the weights are random in the limit when the observational estimator is unbiased.

Proposition 3 (Asymptotic weights: zero bias)**.** *Let the sample sizes be directly proportional[3] ($T = cN$ for some constant $c > 0$) and let the observational data be exogenous ($\boldsymbol{\gamma}_T = 0 \,\forall\, T$). Then, under Assumptions 1 and 2 the estimated weights are random in the limit, i.e.*

$$\lambda^* \longrightarrow 1 - \frac{\operatorname{tr}(\boldsymbol{\Sigma}_0)}{\operatorname{tr}(\boldsymbol{\Sigma}_0) + \frac{1}{c}\operatorname{tr}(\boldsymbol{\Phi}_0)}, \tag{9}$$

$$\hat{\lambda}^{SL} \xrightarrow{d} 1 - \frac{\operatorname{tr}(\boldsymbol{\Sigma}_0)}{\|\zeta\|^2}, \tag{10}$$

$$\hat{\lambda}^{C} \xrightarrow{d} 1 - \frac{\operatorname{tr}(\boldsymbol{\Sigma}_0)}{\operatorname{tr}(\boldsymbol{\Sigma}_0) + \frac{1}{c}\operatorname{tr}(\boldsymbol{\Phi}_0) + \|\zeta\|^2} \tag{11}$$

$$\text{where } \zeta \sim \mathcal{N}\left(0, \boldsymbol{\Sigma}_0 + \frac{1}{c}\boldsymbol{\Phi}_0\right) \tag{12}$$

as $N \to \infty$.

[3] We consider the case where the sample sizes both diverge to infinity ($T = N \to \infty$) together, since then both estimators have non-zero weight rather than one being excluded asymptotically. Furthermore, in this case the limiting weights have the most intuitive interpretation, as the optimal approach is to weigh the estimators proportionally to their limiting variances. Whenever T grows faster than N, the limiting term $\operatorname{tr}(\boldsymbol{\Phi}_0)$ will be eliminated from the (optimal) weight(s), and analogous derivations and observations can be made.

Note that the random weights do not affect the consistency of the weighted estimator (the case $\gamma_0 = 0$ is included in Proposition 1). These random weights in the limit expression of COSE do however give rise to a non-standard asymptotic distribution. Therefore no analytical expression can be derived, so we turn to simulations to investigate its characteristics. In Figures 1 and 2 we plot the histograms of the weights and the COSE based on simulated data using a data generating process (DGP) that is further described in Section 4 (with $\rho = 0, \sigma = 1$). The random limiting behaviour of the weights formalised in Proposition 3 is clearly visible. The SL weight has a very nonstandard distribution, with a large peak at 0 and a rather uniform distribution over the rest of the points. This confirms our earlier observation that the SL weight often yields 0 when endogeneity is absent, fully excluding the experimental data. The conservative weight on the other hand seems to follow a more standard distribution, although its majority of the density is above the optimal weight. This is in line with the limiting expression that contains the estimate of the squared bias that converges to a squared mean-zero normal random variable, which introduces an extra near-zero yet strictly positive value. That results in a weight that is more conservative towards the experimental estimator. In practice however, this is desirable, as in applied work the true level of endogeneity is unknown to the researcher. Our weight does not only weigh the data in an efficient way but also takes an unknown yet possible bias into account, whether actually present or not.

Although both weights produce vastly different limiting distributions, this does not translate into a different limiting behaviour of the COSE itself when weighted with both weights, which is attributed to the fact that both datasets are exogenous. The variance of the weighted estimator is smaller than the experimental estimator alone, even in the limiting case ($N = T = 10,000$) since the weights ensure both datasets are included. This variance reduction is most prominently visible in the MSE plots in Figure 3 that we consider in Section 4.

In summary, these propositions show the self-regulatory behaviour of the weights. When a bias in the observational data is present, COSE converges to the trustworthy and efficient experimental estimator when the experimental sample size grows. When there is no bias, it is beneficial to include both estimators for a minimum overall variance, which is ensured by the random asymptotic weights. Consistency for the true parameter holds for all cases (endogeneity, no endogeneity, for all possible relationships between the two sample sizes), meaning that the COSE will converge to the true value when sample sizes increase.

3.3 Exogeneity Test

We introduce the instrument-free exogeneity test that can be used in our framework to formally test for exogeneity of the observational sample. Classic exogeneity tests require valid instruments, such as it is the case for the Durbin-Wu-Hausman test (Durbin 1954, Wu 1973, Hausman 1978), in which the difference between OLS and IV estimators is considered. In our setting we do not need instrumental variables, but we make use of the exogenous nature of the experimental data instead. Under the null hypothesis of exogeneity, both estimators are unbiased. This ensures that the pivotal statistic as proposed in Proposition 4, has a Chi-squared distribution under the null hypothesis as $N, T \to \infty$.

Figure 1: Distributions of $\lambda^*, \hat{\lambda}^{SL}, \hat{\lambda}^{C}$ for exogenous observational data.

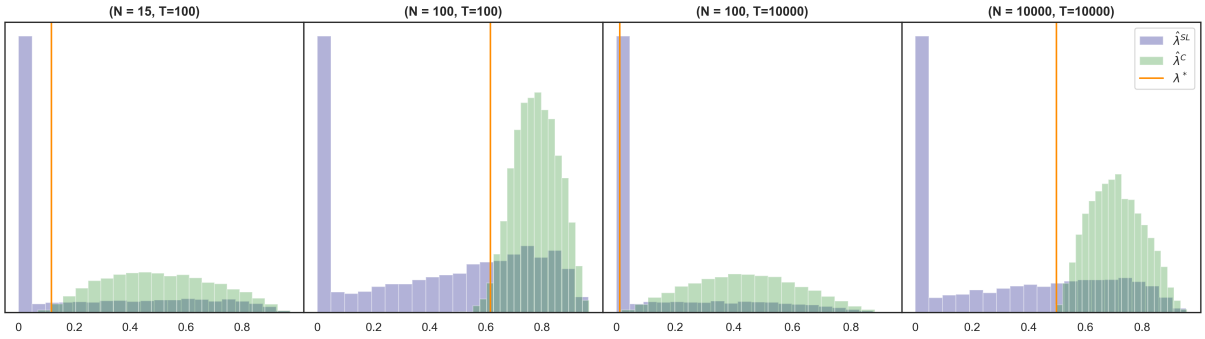
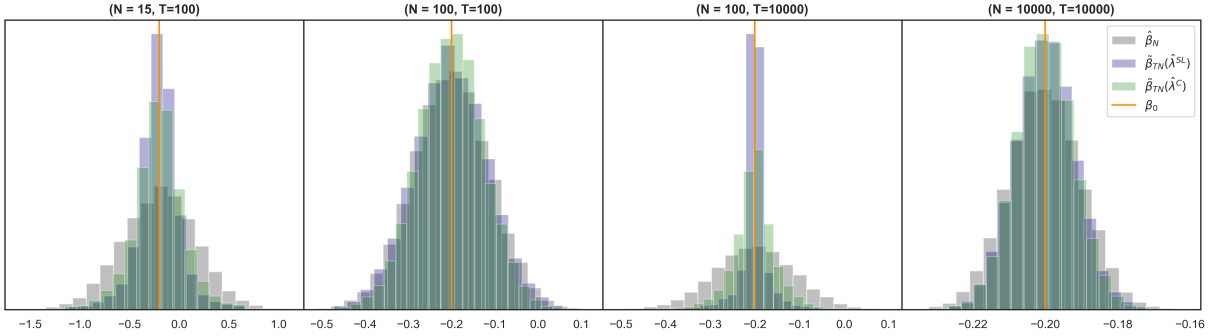


Figure 2: Distributions of $\hat{\boldsymbol{\beta}}_N$, $\tilde{\boldsymbol{\beta}}_{NT}(\hat{\lambda}^{SL})$ and $\tilde{\boldsymbol{\beta}}_{NT}(\hat{\lambda}^{C})$ for exogenous observational data.



Proposition 4 (Exogeneity Test)**.** *Let $T = cN^{\alpha}$, for $c \in \mathbb{N}, \alpha > 0$. Then, under the null hypothesis of exogeneity,*

$$\left(\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_N\right)' \left[\hat{\boldsymbol{\Phi}}_T + \hat{\boldsymbol{\Sigma}}_N\right]^{-1} \left(\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_N\right) \xrightarrow{d} \chi^2_k \qquad \text{as } N \to \infty. \tag{13}$$

Whereas under the alternative the test statistic

$$\left(\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_N\right)' \left[\hat{\boldsymbol{\Phi}}_T + \hat{\boldsymbol{\Sigma}}_N\right]^{-1} \left(\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_N\right) \longrightarrow \infty \qquad \text{as } N \to \infty. \tag{14}$$

Note that the result holds irrespective of how fast T grows with respect to N, as it holds for all $\alpha > 0$.[4]

[4]We can also consider the case where $N, T \in \mathbb{N}$ (corresponding to $\alpha = 0$), but then we need to assume Gausian errors so that the estimators are also normally distributed in finite sample.

Under the alternative, the asymptotic bias in $\hat{\boldsymbol{\beta}}_T$ ensures that the statistic explodes. The power and size of this test will be investigated in Section 4.2.

4 Simulation Study

In this section we show the superiority of COSE in terms of quadratic risk and we investigate the performance of the exogeneity test. Without further specifying the source of endogeneity (omitted variables, measurement error or simultaneity), we simulate the endogeneous data where the first of $k = 5$ regressors, that is $\boldsymbol{x}_{1T}$, is correlated with the error term, while the rest is exogenous and independent of all others. We increase the level of endogeneity by varying correlation ρ between the first regressor and the error term from $\rho = 0$ (no endogeneity) to $\rho = 0.9$ (extremely correlated). We generate

$$\boldsymbol{X}_T = \begin{bmatrix} \boldsymbol{\iota}_T & \boldsymbol{x}_{1T} & \boldsymbol{x}_{2T} & \boldsymbol{x}_{3T} & \boldsymbol{x}_{4T} \end{bmatrix} \tag{15}$$

$$\begin{bmatrix} x_{i,1_T} \\ x_{i,2_T} \\ x_{i,3_T} \\ x_{i,4_T} \\ \varepsilon_{iT} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\iota}_4 \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{I}_4 & \boldsymbol{c}(\sigma,\rho) \\ \boldsymbol{c}(\sigma,\rho)' & \sigma^2 \end{bmatrix} \right) \tag{16}$$

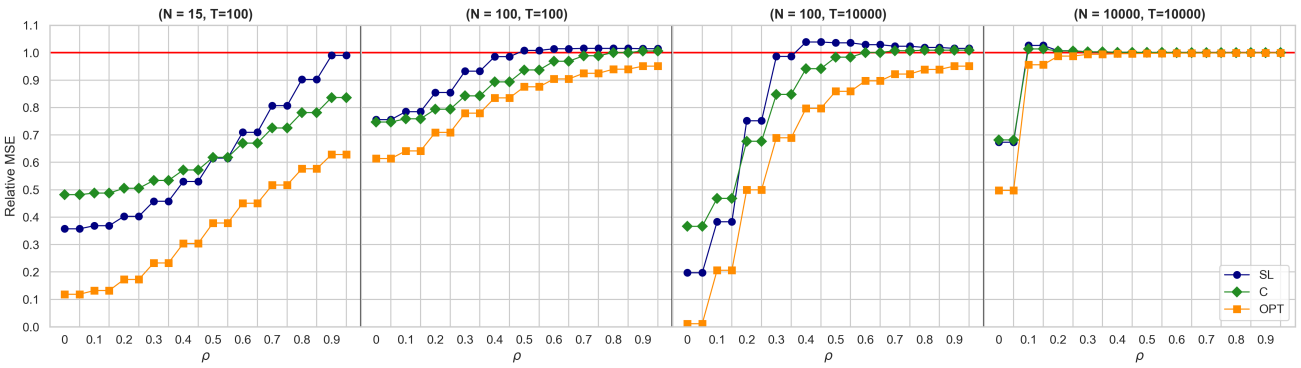
for $i \in \{1, \ldots, T\}$, where $\boldsymbol{c}(\sigma, \rho) = (\sigma\rho, 0, 0, 0)'$. The experimental data matrix $\boldsymbol{X}_N$ is simulated in similar fashion, but with $\rho = 0$.

4.1 MSE Performance Relative to Experimental Estimator

Taking $\boldsymbol{\beta}_0 = (-.3, -.2, .2, -.1, .1)'$, we estimate the OLS estimators $\hat{\boldsymbol{\beta}}_N, \hat{\boldsymbol{\beta}}_T$ and the weighted estimators $\tilde{\boldsymbol{\beta}}_{NT}(\hat{\lambda})$ for different weights $\hat{\lambda}^{SL}$, $\hat{\lambda}^{C}$, λ^* are evaluated for $M = 10,000$ simulations. In Figure 3 we depict the MSE of the weighted estimator relative to the MSE of $\hat{\boldsymbol{\beta}}_N$, the directly available alternative experimental estimator. On the x-axis we put various values of ρ that control the level of endogeneity and for each of these, we have different values of the variance of the error term $\sigma = \sigma_{\varepsilon_N} = \sigma_{\varepsilon_T}$ where $\sigma \in \{1, 10\}$. In Figure 4 we present the average weights along with a box and whisker plot to show the distribution of the simulated values (medians in red). The top row is for $\hat{\lambda}^{SL}$, the bottom for $\hat{\lambda}^{C}$. We only concentrate on the cases where T is larger than or equal to N, to reflect the small experimental sample situation often encountered in real-life settings.

From Figure 3 we learn that with the theoretically optimal weight λ^* (OPT) we can make very large MSE improvements and the weighted estimator never performs worse than the experimental estimator. On the other hand, the estimated weights $\hat{\lambda}$ bring in some variance and finite sample bias, which leads to a higher MSE, but is in the majority of the cases still an improvement over the experimental estimator. The error variance does not have a visible influence on the MSE performance as is evident

Figure 3: MSE of COSE relative to $\hat{\beta}_N$ for different values of endogeneity (ρ). Each value of ρ has two dots representing different levels of standard deviation of the error, for $\sigma = 1$ and $\sigma = 10$ respectively.



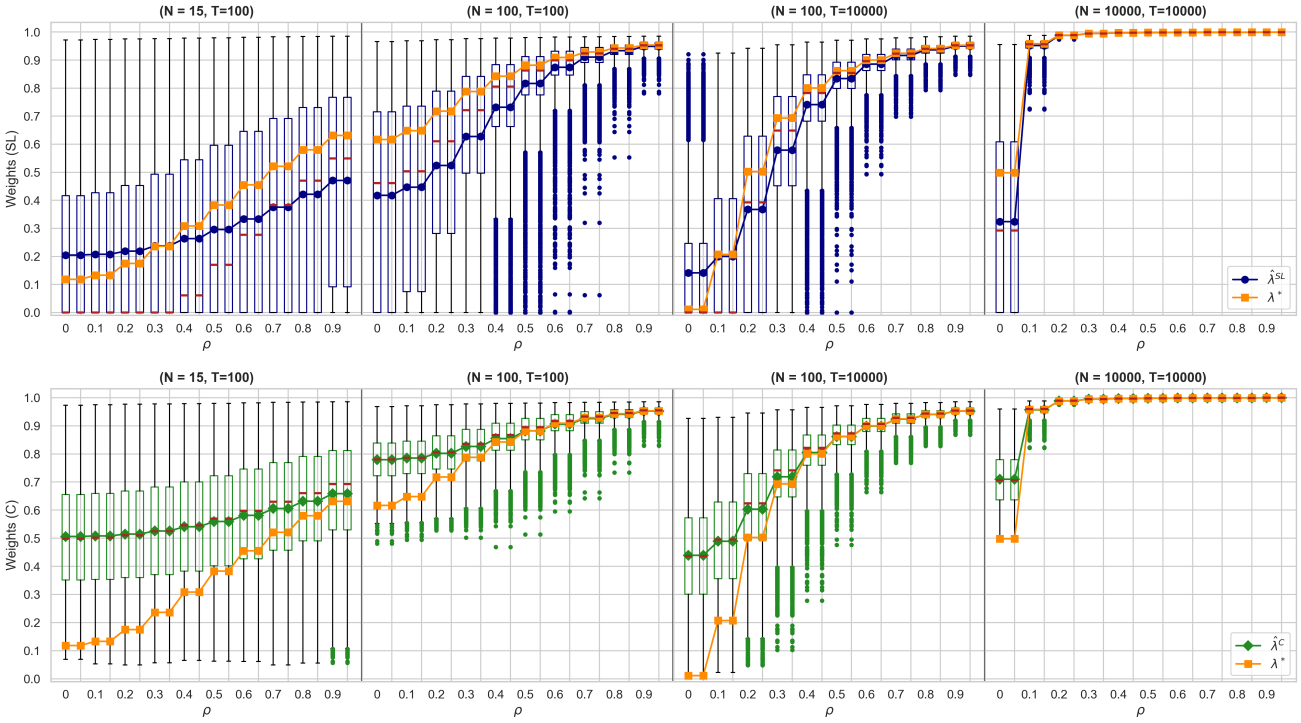
from the step-like course of the curve. In general, the higher the level of endogeneity, the smaller the improvements get, and the larger the observational data set, the higher the gain for lower levels of endogeneity (up to a 60% decrease relative to the experimental estimator for $\rho = 0.1$). The more conservative behaviour of the conservative weight is visible through the fact that it does not reach the possible improvements of the SL weight for lower levels of endogeneity, but simultaneously does not produce a MSE much larger than that of the experimental estimator, in contrast to the SL weight.

The self-regulatory adjustment of the weight is clearly visible in Figure 4. The higher the level of endogeneity, the less weight is put on the observational estimator. Also as variances (relatively) decrease, for example when T grows, the weights adapt. The same re-adjustment is apparent when N increases, with the unbiased estimator becoming more precise while the weight gets closer to 1 and the possible (optimal) improvements in MSE decline.

Between the weights $\hat{\lambda}^{SL}$ and $\hat{\lambda}^{C}$, the conservative weight is visibly more stable judging by the box plots. The range of the conservative weights evaluated on the simulated data is much smaller, while the SL weight often attains very different values when evaluated on different simulated data sets with the same underlying data generating process. Although the average of the SL weight is in many cases closer to the optimal weight, it is often composed of many zero's (median is near zero) and outlying values near one. This is due to its randomness in the denominator that occurs when the estimators are close together which is most evident for low endogeneity ($\rho = 0.1$). The resulting weight becomes highly negative and is then cut off at 0, meaning all weight is put on the efficient observational estimator.

In practice, we will not know whether the two estimators being close together was produced by a coincidental draw of the sampling distribution or by a weak level of endogeneity. For example, for $N = 15, T = 100, \rho = 0.3$, half of the times the SL weight is zero, while there is a considerable amount of endogeneity. Consequently, all weight is put on the observational estimator while an

Figure 4: Boxplots of estimated weights



unbiased estimator is available. Therefore, the more stable behaviour of the conservative weight is desirable in practice. It might overestimate the weight on the experimental data (around 0.52 rather than the optimal 0.23), but it will not fully eliminate the experimental estimator. Ultimately in practical applications it is more intuitive to include experimental data rather than throw it all away. Besides, note that in this simulation study we observe its performance while knowing the exact level of endogeneity ρ. In practice it is unknown, which promotes the use of a conservative approach that harnesses against possible misjudgements even more.

To sum up these observations, the conservative weight is less sensitive than the SL weight, producing more reliable results in a one-off estimation. The SL approach often underestimates the weight, or fully excludes the experimental data, while the conservative approach is more cautious in that respect. Besides, MSE performance is comparable if not favorable for the conservative approach. The SL weight might attain a lower MSE for low levels of endogeneity, but in practice the level of endogeneity is unknown and we cannot truly know whether a case at hand is such a setting in which SL provides the best performance on average.

4.2 Power and Size Exogeneity Test

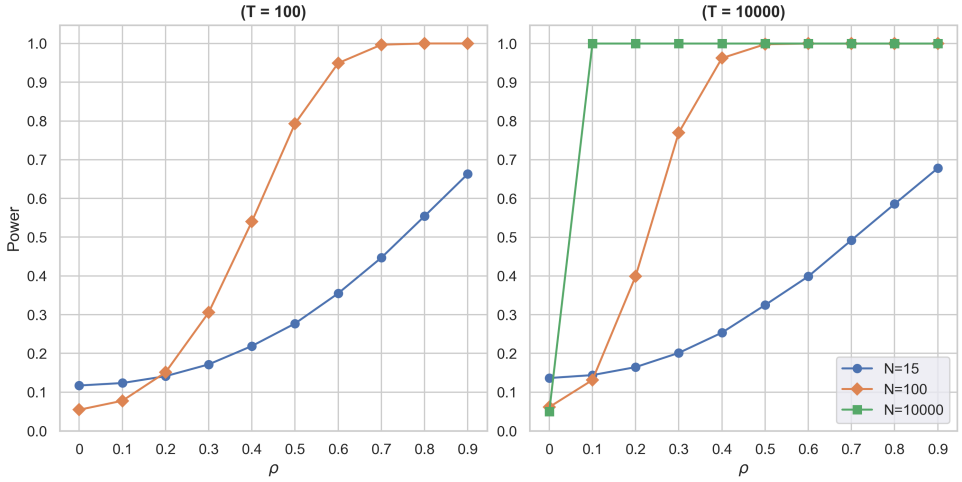
We now analyse the power and size of the instrument-free exogeneity test. Table 1 contains the size of the test for $\sigma = 1$ and $\rho = 0$ (observational data is exogenous) at a significance level of $\alpha = 5\%$, where the data was generated with the linear DGP as described in Section 4.1. The proposed Chi-squared test is slightly too strong for small N, but has appropriate size once N increases.

Table 1: Size exogeneity test ($\alpha = 5\%$)

	χ^2	
	$T = 100$	$T = 10\,000$
$N = 15$	0.117	0.136
$N = 100$	0.054	0.061
$N = 10\,000$	0.062	0.050

Figure 5 shows the power of the exogeneity test for various samples sizes, as ρ increases under the null hypothesis of exogeneity. Power increases are most prominent when both T and N increase, as then a small difference between estimators explodes when scaled by small variances leading to a large value of the test statistic. Low levels of endogeneity are only picked up once N is large. However, a common situation would be one where $T = 10000, N = 100$, in which at $\rho = 0.4$ the rejection rate is already 0.96. Hence, we conclude from these power plots that in our large-T-small-N setting, a non-rejection is most likely indicating that endogeneity is non-existent or low at most.

Figure 5: Power of the Exogeneity Test



5 Empirical Application: Loyalty Discount

We illustrate our method in an empirical study based on a dataset of the Dutch phone repair service company ThePhoneLab, that is interested in identifying the effect of a discount on their customer's purchase behaviour. For a long time only frequent users received a discount, as every year customers that used the service in the past year were sent discount codes. This qualification procedure gives rise to endogeneity, as customers (possibly unknowingly) are self-selected into treatment. The "loyal" group of active customers potentially consists of mostly "clumsy" customers, individuals that value technology more highly, and people with other unobserved characteristics which might lead to higher average sales in the first place, while their frequent visits simultaneously lead to more discounts. The correlation between receiving the discount and these unobserved characteristics leads to a bias in the measurement of the discount effect on sales by comparing the outcomes between both groups directly. To circumvent a misleading result, an A/B test was executed in which only loyal customers were considered. Some were randomly excluded from receiving the discount, allowing for reliable measurement of the discount effect.

Several studies have investigated the effects of promotions on sales, for example whether long-term effects differ between frequent and occasional customers (Lim et al. (2005), Reimer et al. (2014)). In these studies the most sensitive group to price promotions differs across different types of products (perishable, non-perishable goods, digital music products) and ThePhoneLab would like to investigate the effect of discounts on their repair service based on the short-term experiment they executed. In particular, we measure the effect of a discount on both net sales and number of sold products for the average customer. Given that the number of experimental observations is much smaller than the observational ones, we show that experimentation alone is not enough to measure a significant effect, and by including the observational data we provide a more accurate measurement of said effect.

5.1 Empirical Model

The empirical model aims to describe sales (s_i) as a function of the discount D_i (1 if received, 0 if otherwise), additional observed control variables $\boldsymbol{X}_i$, and true loyalty L_i (1 if truly loyal, 0 otherwise), that is

$$s_i = \beta_0 + \beta_1 D_i + \boldsymbol{\beta}_2 \boldsymbol{X}_i + \beta_3 L_i + u_i, \tag{17}$$

for $i = 1, \ldots, T$, where β_j is a regression coefficients ($j = 0, \ldots, 3$) and u_i is the normally distributed error term. Our prime interest is β_1 which measures the effect of providing the discount[5]. The true

[5]Note that we are interested in the effect of the discount on the average customer, not on loyal people. Using experimentation, this can clearly be done by giving random customers (loyal and non-loyal) a discount and evaluating the Average Treatment Effect (ATE). However, for the company in question the only available record of giving out any discounts is by providing them only to (part of the) loyal customers. Thus, to make claims about the effect on the average customer in the experimental data set which is comprised of loyal customers, we subject to the implicit assumption that the effect is the same for loyal and non-loyal customers.

loyalty term L_i represents all unobserved characteristics that contribute to the frequent use of the service. The issue here is that next to L_i being unobserved, during the observational period it is correlated with D_i, leading to a omitted variable bias in the estimate of β_1. If we could find a proxy for loyalty, denoted $\tilde{L}_i$, by flagging a customer as loyal based on their purchase behaviour, we could control for it. However, such a variable is likely to be perfectly correlated with the treatment in the observational data as loyal customers were given the discount, leaving β_1 unidentified. Note that during the experimental period, treatment is uncorrelated with L_i, as among those that were entitled to the discount, a certain percentage was randomly excluded from actually receiving it. In this case, omitting L_i will not lead to such a bias and we get a consistent estimate for β_1.

In many settings, perhaps more information would be available that could be exploited to estimate the causal effect in the observational data using quasi-experimental methods. For example, if the discounts were given after a certain variable (e.g. number of purchases) crossed a threshold, regression discontinuity design could be applied. Although in this case the actual treatment was based on some decision rule, in this dataset there is no such threshold variable that consistently determines treatment throughout the whole time period. Alternatively, one could consider the difference-in-differences estimator if the discounts were given out at the same time for all treated individuals. In the application at hand the discounts were given at different points in time, although evenly spaced over the interval, hence we do not know the pre-treatment period of the untreated. Finally, if valid instruments were be available, a two-stage-least-squares estimator could give an unbiased estimate of the causal effect. In absence of any such additional variables or design characteristics, we simply continue with an OLS estimate that most likely suffers from the omitted variable bias as our observational estimator. However, even if such a quasi-experimental method would be applicable, we would use the quasi-experimental estimate instead, as there always remains some uncertainty about whether the appropriate assumptions (e.g. parallel trends, exclusion restriction) are satisfied and thus whether the estimator is truly unbiased. If the assumptions are in fact satisfied, we deal with two unbiased estimators, which our simulation study shows will still lead to a lower MSE.

5.2 Data

For each individual customer in the dataset we accumulate total net sales (sales minus discounts) in Euros and number of products sold within both observational and experimental periods, for those who received the discount (`discount_given`=1) and those who did not receive the discount (`discount_given`=0). In the experimental data we only accumulate all post-treatment sales, since we know the exact date when an individual received the discount (treatment) or was actively excluded (non-treatment). In the observational data, we do not observe such a pre- or post-treatment date, so we aggregate sales for each individual over the entire observational time period. Other available characteristics are `gender` and `ZIP` code. The latter allows the creation of an urbanity index (1= urban, 5 = rural) and average home property value in the area. In the observational sample around 11% of the customers were flagged as loyal and all received the discount, whereas in the experimental sample all were flagged as loyal, but 68% of them eventually received the discount. The observational sample size

is considerably larger than the experimental sample size ($T = 14,236$ and $N = 110$), mostly due to the fact that in the experimental data set we only consider the loyal group, while in the observational data all customers during that period are taken into account. Furthermore, the experiment was only run for a couple of months, while the observational data was accumulated over many years.

5.3 Results

In Table 2 we present the estimation results of OLS regressions applied to the empirical model in equation (17) (where L_i is omitted) using the observational and experimental data, for two different dependent variables. In the first column we observe that the observational data suggests that the discount provides incentives to customers to buy more products as the coefficient is highly significant and positive, although relatively small. We do however suspect that this result is biased since customers self-select into receiving the discount during this observational period. To formally investigate this, we evaluate the exogeneity test and observe in Table 3 that we can indeed reject the null hypothesis of exogeneity of the observational data. We therefore turn to a more reliable estimator, and discover that the experimental results give a smaller and non-significant effect, suggesting that the discount may not encourage customers to purchase more products. This non-signficance could be a result of the near-zero coefficient, but the large increase in variance due to the small sample size also plays a role, which manifests in large standard errors and a non-rejection for the F-test of overall significance.

To investigate if the discount really does not affect the quantity of products sold even when the influence of the large variance is mitigated, we consult our COSE estimates in Table 4. We include the estimated weight, the weighted COSE coefficients with both SL and consistent weights and bootstrap confidence intervals, alongside the original experimental estimator. Considering the coefficient of `discount_given`, we still have an insignificant effect of the discount after weighing the observational and experimental results with weights accounting for the apparent bias and variances that are present. The estimated SL and consistent weights are similar (0.51 and 0.67), both putting most weight on the experimental estimator. This is in line with the result of the exogeneity test that is strongly rejected, as one should be less comfortable with putting a lot of weight on the observational estimator, given it is suspected to be biased. It leads to an insignificant COSE coefficient in agreement with the experimental estimator.

We established that the discount does not stimulate customers to buy more products. It is expected that the issuing of discount codes is harmful to net sales as it costs more than it yields. This suspicion appears to be confirmed by the OLS results for `total_net_sales` in Table 2. The observational data yields a highly significant negative value, implying a negative effect of the discount on net sales. However, the rejection[6] of the exogeneity test in Table 3 suggests we should not trust these observational results. Consulting the experimental data instead, we actually find an even larger negative coefficient of the discount, indicating that the observational data understates the effect on

[6]Since a rejection would indicate the presence of endogeneity, we recommend taking a high significance level is a conservative approach that is sceptical of the exogenous nature of observational data.

Table 2: OLS regression results loyalty discount for observational and experimental samples.

Dependent variable	number_of_products		total_net_sales	
	Obs.	Exp.	Obs.	Exp.
constant	1.184^{***}	1.026^{***}	107.120^{***}	122.156^{***}
	(0.013)	(0.082)	(2.665)	(28.423)
discount_given	0.044^{***}	0.016	-7.621^{***}	-20.852
	(0.012)	(0.053)	(2.220)	(17.402)
gender_female	-0.019^{**}	0.004	2.000	-0.915
	(0.008)	(0.051)	(1.410)	(16.608)
urbanity_index	-0.009^{***}	0.004	-1.331^{***}	4.424
	(0.003)	(0.029)	(0.507)	(9.866)
property_valuation	0.000^{***}	0.000	0.011^{*}	0.0393
	(0.000)	(0.000)	(0.006)	(0.045)
% discount_given	11.5 %	68.8 %	11%	68.2%
sample size[1]	$T = 15\,195$	$N = 151$	$T = 14\,236$	$N = 110$
F-test joint significance (p-val)	0.00^{***}	0.869	0.00^{***}	0.709

Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

[1]Sample sizes differ for different dependent variables due to outlier removal.

Table 3: Exogeneity test for loyalty discount data

	χ^2 statistic	p-value
number_of_products	29.182	0.0000^{***}
total_net_sales	10.924	0.0529^{*}

net sales. However, we cannot make such a claim with certainty since the experimental result is insignificant. In this case, the insignificance can most likely be attributed to the small sample size, as the standard errors are substantial and the joint F-test is again insignificant.

Although we have a strong suspicion that the discount has a negative effect on total net sales, the large variance due to limited experimental data leaves the evidence as inconclusive. To clarify the current ambiguity of the results, we again turn to our method as it compromises between the two sides of the bias-variance trade-off. From the results in Table 4 we can conclude that our estimator indeed provides

Table 4: Weighted estimator regression results. For both regressions we compare the experimental estimator to COSE, reporting estimated weights $\hat{\lambda}^{SL}, \hat{\lambda}^{C}$ and the corresponding COSE coefficient with these estimated weights. We report 95% confidence intervals for this coefficient for COSE obtained with a case bootstrap, while for $\hat{\beta}_N$ the CI is based on its asymptotic distribution. Significance of the coefficients is based on the bootstrapped confidence intervals.

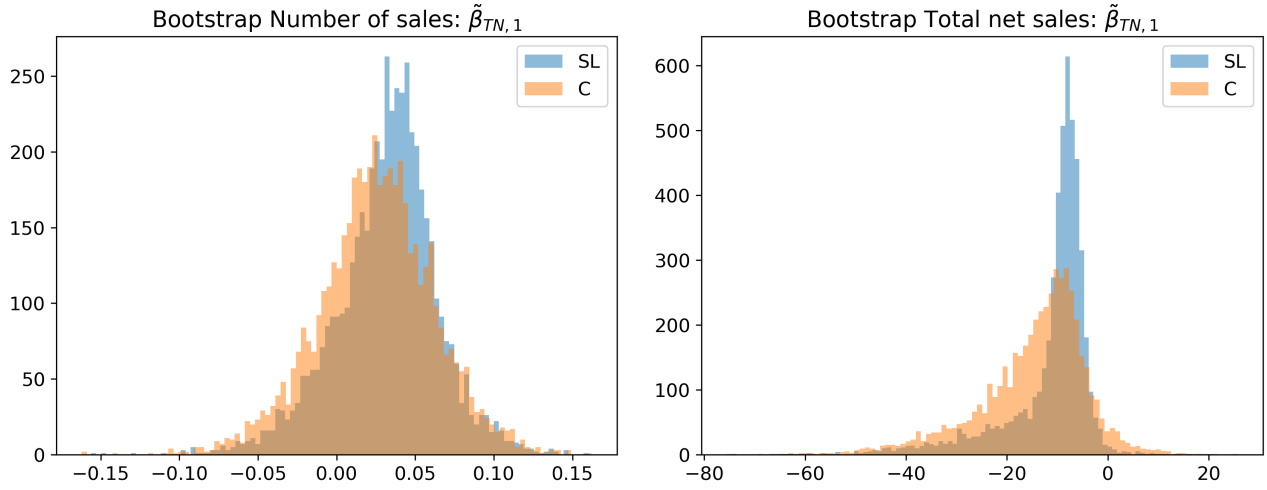
	number_of_products			total_net_sales		
$\tilde{\beta}_{TN}(\hat{\lambda})$	Exp.	COSE(SL)	COSE(C)	Exp.	COSE(SL)	COSE(C)
constant	1.026^{***}	1.104^{***}	1.078^{***}	122.155^{***}	107.120^{***}	110.660^{***}
	[0.86, 1.19]	[0.87, 1.20]	[0.87, 1.20]	[65.8, 178.5]	[95.9, 154.0]	[86.0, 158.1]
discount given	0.0162	0.030	0.025	−20.852	-7.620^{**}	-10.736^{*}
	$[-0.09, 0.12]$	$[-0.04, 0.09]$	$[-0.05, 0.1]$	$[-55.4, 13.6]$	$[-37.8, -2.4]$	$[-41.4, 2.3]$
gender female	0.004	−0.008	−0.004	−0.915	2.000	1.314
	[-0.10, 0.10]	[-0.06, 0.08]	[-0.07, 0.08]	[-33.8, 32.0]	[-15.9, 12.8]	[-20.6, 18.7]
urbanity index	0.004	−0.003	−0.001	4.424	−1.331	0.024
	[-0.05, 0.06]	[-0.03, 0.06]	[-0.04,0.06]	[-15.1, 23.9]	[-4.4, 7.2]	[-6.7, 10.3]
property val	0.000	0.000	0.000	0.039	0.011	0.017
	[-0.00, 0.00]	[-0.00, 0.00]	[-0.00, 0.00]	[-0.05 0.1]	[-0.02, 0.05]	[-0.03, 0.08]
$\hat{\lambda}$	1	0.506	0.672	1	0	0.235

Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

such significant negative coefficients for both types of weights. The COSE estimator tells us there is in fact a significant cost to be paid by awarding discounts to customers. We have found that giving out discounts does not lead to a significant increase in the amount of products a customer buys, which is the main driver of this negative impact on net sales. A word of caution is appropriate however when it comes to the significance of the estimates, that are evaluated using a bootstrap procedure presented is analyzed the Appendix. The weighted estimator has a finite sample bias through the introduction of the biased observational data, which potentially manipulates A significance finding. Nevertheless, compared to the unbiased experimental estimator, the observational estimator turns out to be biased towards zero. So, although the COSE is a biased estimator, it is not the reason for its significance. In fact, *despite* its bias towards zero it remains significant as it takes over a smaller variance from the observational estimator. Accounting for this bias would probably lead to a higher level of significance. Therefore we do not suppose it is likely that this finding is incidental. In Appendix B we show that the bootstrap procedure provides a good estimate of the finite sample distribution by performing a Monte Carlo study based on the original data.

Moreover, this application clearly illustrates the advantage of the conservative weight over the SL weight. We obtain an estimated weight $\hat{\lambda}^{SL} = 0$, meaning the experimental data is fully excluded, as a result of a huge variance and a relatively small difference between the two estimators which results in a negative weight (-2.35) if not truncated at 0. Excluding the experimental data and fully relying on the observational result is quite a bold measure, since the exogeneity test indicates that the observational data is likely to be endogenous. The zero weight results in a bootstrapped confidence interval that is based on the small variance observational estimator alone, resulting in almost half the volume (51% for `discount_given`) of the experimental confidence interval, and leads to a significant result. The conservative weight however does take the experimental data into account with $\hat{\lambda}^{C} = 0.23$. This gives a slightly wider confidence interval (a 37% decrease in volume for `discount_given`) but still leads to a significantly negative result. This is clearly apparent in Figure 6 that contains the distribution of the bootstrapped weighted estimator. This distribution is clearly asymmetric with mostly negative support.

Figure 6: Bootstrap distributions of COSE with SL and C weights. These bootstrap histograms are for the weighted estimates of `discount_given` ($\tilde{\beta}_{TN,1}(\hat{\lambda})$) in particular.



This empirical study highlights both the drawbacks of experimentation and the strengths of our method. With observational data one might yield highly significant intuitive results, but this does not measure the true effect of the intervention by the company as it understates the actual effect in this case. At the same time, the small sample size of the experimental data leads to larger variances of the estimates that drown out the significance of the effect, leading to an ambiguous outcome. But our method shows that combining the already available experimental and otherwise useless observational data can help to circumvent this problem. Based on our proposed weighted estimator, we have been able to conclude that the discount has a significant negative effect on total net sales.

6 Conclusion

In this study we have reviewed a useful method for practitioners for combining observational and experimental data to uncover causal estimates with more precision than the experimental estimator itself. We have shown in a simulation study that large MSE improvements can be made when including the observational data with a self-regulatory weight that balances the bias and variance appropriately. Our proposed alternative conservative weight keeps both experimental and observational data sets active which is useful in the one-off estimation exercises that are done in practice. Furthermore, we have introduced an exogeneity test that can determine whether one is indeed dealing with endogenous data. In an application to loyalty discounts we have shown its advantages. Our proposed weighted estimator was conclusive whereas experimental data alone gave rise to ambiguous outcomes due to the small sample size. Using our method we have established a significant negative effect of the discount on total net sales, while the number of sales have not been significantly affected by the discount.

Our approach is straightforward to implement for practitioners, without the need for a exhaustive understanding of statistics. It is crucial that marketing researchers and data analysts are equipped with accessible, understandable and inexpensive methods that address endogeneity, since it is an intricate issue and large scale marketing decisions might be based on the modelling of data relations that suffer from this phenomenon. However, fully eliminating endogeneity while also collecting large enough sample sizes is a great challenge. Taking these practical considerations in mind, we provide a solution that reduces overall estimation risk.

Our proposed method is more widely applicable than we have set out in this study. Although we set off with a linear model with normally distributed errors, the method can straightforwardly be extended with its relience on other estimators (nonlinear least squares, logistic, maximum likelihood estimators for different distributions). One could even consider combining different types of estimators on each dataset, or replacing the experimental estimator by an IV estimator when strong instruments are available. Eventually, the weight only requires the estimators themselves together with consistent estimates of the variance matrices. However, the same MSE improvements cannot always be guaranteed for nonlinear extensions in particular, since the variance matrix often needs to be evaluated at the estimator, which brings about extra variance in the estimated weight and the resulting COSE, making it harder to be a competitive alternative to the experimental estimator. Nevertheless, there will always be MSE reductions for the optimal weight whichever estimators are employed.

References

Anderson TW (1984) *An introduction to multivariate statistical analysis* (New York: John Wiley & Sons), 2nd edition.

Athey S, Chetty R, Imbens G (2020) Combining experimental and observational data to estimate treatment effects on long term outcomes. Preprint, submitted June 17, https://arxiv.org/abs/2006.09676.

Campbell C, Runge J, Bates K, Haefele S, Jayaraman N (2022) It's time to close the experimentation gap in advertising: Confronting myths surrounding ad testing. *Business Horizons* 65(4):437–446.

Cooper GF, Yoo C (2013) Causal discovery from a mixture of experimental and observational data. Preprint, submitted January 23, https://arxiv.org/abs/1301.6686.

Cui R, Zhang DJ, Bassamboo A (2019) Learning from inventory availability information: Evidence from field experiments on amazon. *Management Science* 65(3):1216–1235.

Durbin J (1954) Errors in variables. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute* 22(1/3):23–32.

Ebbes P, Papies D, van Heerde HJ (2022) Dealing with endogeneity: A nontechnical guide for marketing researchers. Homburg C, Klarmann M, Vomberg A, eds., *Handbook of Market Research*, 181–217 (Cham: Springer).

Fernandez Loria CM, Provost F (2020) Combining observational and experimental data to improve large-scale decision making. *ICIS 2020 Proceedings.*

Gasse M, Grasset D, Gaudron G, Oudeyer PY (2021) Causal reinforcement learning using observational and interventional data. Preprint, submitted June 28, https://arxiv.org/abs/2106.14421.

Gluck M (2011) Best practices for conducting online ad effectiveness research. Report, Interactive Advertising Bureau, New York.

Gordon BR, Zettelmeyer F, Bhargava N, Chapsky D (2019) A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science* 38(2):193–225.

Green EJ, Strawderman WE (1991) A james-stein type estimator for combining unbiased and possibly biased estimators. *Journal of the American Statistical Association* 86(416):1001–1006.

Gui GZ (2024) Combining observational and experimental data to improve efficiency using imperfect instruments. *Marketing Science* 43(2):378–391.

Hausman JA (1978) Specification tests in econometrics. *Econometrica* 46(6):1251–1271.

James W, Stein C (1961) Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* 1:361–380.

Judge GG, Mittelhammer RC (2004) A semiparametric basis for combining estimation problems under quadratic loss. *Journal of the American Statistical Association* 99(466):479–487.

Kohavi R, Longbotham R, Sommerfield D, Henne RM (2009) Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery* 18:140–181.

Kohavi R, Tang D, Xu Y (2020) *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing* (Cambridge: Cambridge University Press).

Lewis RA, Rao JM, Reiley DH (2011) Here, there, and everywhere: correlated online behaviors can lead to overestimates of the effects of advertising. *Proceedings of the 20th international conference on World wide web*, 157–166.

Li L, Chen S, Kleban J, Gupta A (2015) Counterfactual estimation and optimization of click metrics in search engines: A case study. *Proceedings of the 24th International Conference on World Wide Web*, 929–934.

Lim J, Currim IS, Andrews RL (2005) Consumer heterogeneity in the longer-term effects of price promotions. *International Journal of Research in Marketing* 22(4):441–457, ISSN 0167-8116, URL http://dx.doi.org/https://doi.org/10.1016/j.ijresmar.2005.09.006.

Mittelhammer RC, Judge GG (2005) Combining estimators to improve structural model estimation and inference under quadratic loss. *Journal of Econometrics* 128(1):1–29, ISSN 0304-4076, URL http://dx.doi.org/https://doi.org/10.1016/j.jeconom.2004.08.006.

Nunan D, Di Domenico M (2022) Value creation in an algorithmic world: towards an ethics of dynamic pricing. *Journal of Business Research* 150:451–460.

Papies D, Ebbes P, Van Heerde HJ (2017) Addressing endogeneity in marketing models. Leeflang PSH, Wieringa JE, Bijmolt TH, Pauwels KH, eds., *Advanced Methods for Modeling Markets*, 581–627 (Cham: Springer International Publishing).

Reimer K, Rutz OJ, Pauwels K (2014) How online consumer segments differ in long-term marketing effectiveness. *Journal of Interactive Marketing* 28(4):271–284, URL http://dx.doi.org/10.1016/j.intmar.2014.05.002.

Rosenman ET, Basse G, Owen AB, Baiocchi M (2023) Combining observational and experimental datasets using shrinkage estimators. *Biometrics* 00:1–13.

Rutz OJ, Watson GF (2019) Endogeneity and marketing strategy research: An overview. *Journal of the Academy of Marketing Science* 47:479–498.

Stein C (1956) Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, 197–207 (Berkeley: University of California Press).

Wu DM (1973) Alternative tests of independence between stochastic regressors and disturbances. *Econometrica* 41(4):733–750.

Appendix A: Proofs

Proof of Proposition 1

We write the two estimators as

$$\tilde{\boldsymbol{\beta}}_{TN}(\hat{\lambda}^{SL}) = \hat{\boldsymbol{\beta}}_N + \frac{\operatorname{tr}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{N}})}{\left\|\hat{\boldsymbol{\beta}}_N - \hat{\boldsymbol{\beta}}_T\right\|^2}(\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_N)$$

$$\tilde{\boldsymbol{\beta}}_{TN}(\hat{\lambda}^{C}) = \hat{\boldsymbol{\beta}}_N + \frac{\operatorname{tr}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{N}})}{\operatorname{tr}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{N}}) + \operatorname{tr}(\hat{\boldsymbol{\Phi}}_{\boldsymbol{T}}) + \left\|\hat{\boldsymbol{\beta}}_N - \hat{\boldsymbol{\beta}}_T\right\|^2}(\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_N).$$

First consider the case of a nonzero limiting bias ($\boldsymbol{\gamma}_0 \neq 0$) and an increasing sample size T ($\alpha > 0$). Then,

$$\operatorname*{plim}_{N\to\infty} \tilde{\boldsymbol{\beta}}_{TN}(\hat{\lambda}^{SL}) = \operatorname*{plim}_{N\to\infty} \tilde{\boldsymbol{\beta}}_{TN}(\hat{\lambda}^{C}) = \operatorname*{plim}_{N\to\infty} \hat{\boldsymbol{\beta}}_N + \frac{0}{\boldsymbol{\gamma}_0'\boldsymbol{\gamma}_0}\boldsymbol{\gamma}_0 = \boldsymbol{\beta}_0.$$

Even when the sample size T does not grow ($\alpha = 0$), we get consistency since $\operatorname*{plim}_{N\to\infty} \tilde{\boldsymbol{\beta}}_{TN}(\hat{\lambda}^{SL}) = \boldsymbol{\beta}_0 + \frac{0}{\boldsymbol{\gamma}_T'\boldsymbol{\gamma}_T}\boldsymbol{\gamma}_T$ and $\operatorname*{plim}_{N\to\infty} \tilde{\boldsymbol{\beta}}_{TN}(\hat{\lambda}^{C}) = \boldsymbol{\beta}_0 + \frac{0}{\operatorname{tr}(\hat{\boldsymbol{\Phi}}_{\boldsymbol{T}})+\boldsymbol{\gamma}_T'\boldsymbol{\gamma}_T}\boldsymbol{\gamma}_T$.

Now consider the case in which the observational estimator is consistent itself ($\boldsymbol{\gamma}_0 = 0$). We can then write

$$\tilde{\boldsymbol{\beta}}_{TN}(\hat{\lambda}^{SL}) = \hat{\boldsymbol{\beta}}_N + \frac{N \operatorname{tr}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{N}})}{\left\| \sqrt{N}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0) - \sqrt{N}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0) \right\|^2} (\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_N).$$

Considering the limiting behaviour of each element using stochastic order notation for $\alpha > 0$, we get $(\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_N) = o_p(1)$ due to the zero limit bias. By Assumptions 1 and 2, $N \operatorname{tr}(\hat{\Sigma}_N) = O_p(1)$,$\sqrt{N}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0) = O_p(1)$ and $\sqrt{N}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0) = N^{(1-\alpha)/2}\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0) = N^{(1-\alpha)/2}O_p(1)$. We then get

$$\operatorname*{plim}_{N\to\infty} \tilde{\boldsymbol{\beta}}_{TN}(\hat{\lambda}^{SL}) = \operatorname*{plim}_{N\to\infty} \hat{\boldsymbol{\beta}}_N + \frac{O_p(1)}{O_p(1) + O_p(N^{(1-\alpha)})} o_p(1) = \boldsymbol{\beta}_0$$

for all $\alpha > 0$. A similar calculation can be done for $\hat{\lambda}^C$, albeit with an extra $O_p(N^{1-\alpha})$ term in the denominator. Also there, the convergence to $\boldsymbol{\beta}_0$ is driven by the vanishing difference between the two estimators and is even accelerated by an exploding denominator when $0 < \alpha < 1$.
In the specific case of $\alpha = 0$, $(\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_N) = O_p(1)$, but consistency is guaranteed for both weights due to a dominating $O_p(N)$ term in the denominator of the fraction.

Proof of Proposition 2
Regardless of α, $\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_N = O_p(1)$. For $\alpha > 0$, $\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_N \xrightarrow{p} \boldsymbol{\gamma}_0$ as $N \to \infty$ with $T = cN^\alpha$, while for fixed $T \in \mathbb{N}$ ($\alpha = 0$), we have $\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_N \underset{N\to\infty}{\simeq} \hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0$ which is $O_p(1)$ by Assumption 2. First considering $\hat{\lambda}^{SL}_*$,

$$\begin{aligned}
\sqrt{N}\left(\tilde{\boldsymbol{\beta}}_{TN}(\hat{\lambda}^{SL}_*) - \boldsymbol{\beta}_0\right) &= \sqrt{N}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0) + \frac{\sqrt{N}\operatorname{tr}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{N}})}{\left\|\hat{\boldsymbol{\beta}}_N - \hat{\boldsymbol{\beta}}_T\right\|^2}(\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_N) \\
&= \sqrt{N}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0) + \frac{\sqrt{N}O_p\left(\frac{1}{N}\right)}{\|O_p(1)\|^2} O_p(1) \\
&= \sqrt{N}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0) + O_p\left(\frac{1}{\sqrt{N}}\right) \xrightarrow{d} \mathcal{D}
\end{aligned}$$

as $N \to \infty$. Similarly, for $\hat{\lambda}^C_*$,

$$\begin{aligned}
\sqrt{N}\left(\tilde{\boldsymbol{\beta}}_{TN}(\hat{\lambda}^{C}_*) - \boldsymbol{\beta}_0\right) &= \sqrt{N}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0) + \frac{\sqrt{N}\operatorname{tr}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{N}})}{\operatorname{tr}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{N}}) + \operatorname{tr}(\hat{\boldsymbol{\Phi}}_{\boldsymbol{T}}) + \left\|\hat{\boldsymbol{\beta}}_N - \hat{\boldsymbol{\beta}}_T\right\|^2}(\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_N) \\
&= \sqrt{N}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0) + \frac{O_p\left(1\right)}{O_p\left(\frac{1}{\sqrt{N}}\right) + O_p\left(\frac{\sqrt{N}}{T}\right) + \sqrt{N}\left\|O_p(1)\right\|^2} O_p(1) \\
&= \sqrt{N}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0) + O_p\left(\frac{1}{\sqrt{N}}\right) \xrightarrow{d} \mathcal{D}
\end{aligned}$$

as $N \to \infty$.

Proof of Proposition 3

Let $T = cN$, $\boldsymbol{\gamma}_T = 0$ and define $\zeta \sim \mathcal{N}(0, \boldsymbol{\Sigma}_0 + \frac{1}{c}\boldsymbol{\Phi}_0)$. When we multiply numerators and denominators with N, we get

$$1 - \lambda^* = \frac{N \operatorname{tr}(\boldsymbol{\Sigma}_N)}{N \operatorname{tr}(\boldsymbol{\Sigma}_N) + \frac{1}{c} T \operatorname{tr}(\boldsymbol{\Phi}_T)} \longrightarrow \frac{\operatorname{tr}(\boldsymbol{\Sigma}_0)}{\operatorname{tr}(\boldsymbol{\Sigma}_0) + \frac{1}{c}\operatorname{tr}(\boldsymbol{\Phi}_0)} \text{ as } N \to \infty$$

$$1 - \hat{\lambda}^{SL} = \frac{N \operatorname{tr}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{N}})}{\left\| \sqrt{N}(\hat{\boldsymbol{\beta}}_N - \hat{\boldsymbol{\beta}}_T) \right\|^2} \xrightarrow{d} \frac{\operatorname{tr}(\boldsymbol{\Sigma}_0)}{\|\zeta\|^2} \text{ as } N \to \infty$$

$$1 - \hat{\lambda}^{C} = \frac{N \operatorname{tr}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{N}})}{N \operatorname{tr}(\hat{\boldsymbol{\Sigma}}_N) + \frac{1}{c} T \operatorname{tr}(\hat{\boldsymbol{\Phi}}_T) + \left\| \sqrt{N}(\hat{\boldsymbol{\beta}}_N - \hat{\boldsymbol{\beta}}_T) \right\|^2} \xrightarrow{d} \frac{\operatorname{tr}(\boldsymbol{\Sigma}_0)}{\operatorname{tr}(\boldsymbol{\Sigma}_0) + \frac{1}{c}\operatorname{tr}(\boldsymbol{\Phi}_0) + \|\zeta\|^2} \text{ as } N \to \infty,$$

using Assumptions 1 and 2 and by recognising the joint convergence in distribution of all terms in the fractions and applying the continuous mapping theorem. Although the rational function $f(x) = \frac{1}{x^2}$ that is discontinuous at zero seems to invalidate this last step, the continuous mapping theorem states that as long as the probability is zero at discontinuous points ($\mathbb{P}[\zeta = 0] = 0$, which is clearly the case with continuous distributions), the result holds.

Proof of Proposition 4

We multiply and divide the test statistic by $\sqrt{NT}$, so that we can use the asymptotic distributions from Assumptions 1 and 2. Under the null hypothesis of exogeneity,

$$\begin{aligned}\left[\hat{\boldsymbol{\Phi}}_T + \hat{\boldsymbol{\Sigma}}_N\right]^{-1/2} \left(\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_N\right) &= \left[NT\hat{\boldsymbol{\Phi}}_T + TN\hat{\boldsymbol{\Sigma}}_N\right]^{-1/2} \left(\sqrt{N}\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}) - \sqrt{T}\sqrt{N}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta})\right) \\ &\xrightarrow{d} \left[N\boldsymbol{\Phi}_0 + T\boldsymbol{\Sigma}_0\right]^{-1/2} \mathcal{N}\left(0, N\boldsymbol{\Phi}_0 + T\boldsymbol{\Sigma}_0\right) \quad \text{as } N \to \infty \\ &\sim \mathcal{N}(\mathbf{0}, I_k)\end{aligned}$$

by Slutsky's Theorem. The test statistic is the inner product of this term, which yields the χ^2 distribution. Under the alternative hypothesis,

$$\left[\hat{\boldsymbol{\Phi}}_T + \hat{\boldsymbol{\Sigma}}_N\right]^{-1/2} \left(\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_N\right) \xrightarrow{d} \left[N\boldsymbol{\Phi}_0 + T\boldsymbol{\Sigma}_0\right]^{-1/2} \mathcal{N}\left(\sqrt{NT}\boldsymbol{\gamma}_0, N\boldsymbol{\Phi}_0 + T\boldsymbol{\Sigma}_0\right) \longrightarrow \infty \quad \text{as } N \to \infty.$$

Appendix B: Bootstrap Performance

Bootstrap Procedure

The bootstrap procedure we implement for the application is rather straightforward. It is based on resampling whole rows of the data matrix, to ensure that the potential endogenous relationship between y and $\boldsymbol{X}$ is maintained.

Step 1. For each bootstrap iteration b, resample from the experimental data N observations with replacement denoted by $y_{i,b}^N$ and $\boldsymbol{X}_{i,b}^N$.
In similar fashion, resample the observational data with replacement.

Step 2. With these experimental and observational bootstrap data sets, estimate the OLS estimators $\hat{\boldsymbol{\beta}}_N^{(b)}$, $\hat{\boldsymbol{\beta}}_T^{(b)}$ and the variances $\hat{\boldsymbol{\Sigma}}_N^{(b)}$ and $\hat{\boldsymbol{\Phi}}_T^{(b)}$.

Step 3. Using the estimated quantities in Step 2, calculate $\hat{\lambda}^{SL\,(b)}$ and $\hat{\lambda}^{C\,(b)}$ and the corresponding COSE $\tilde{\boldsymbol{\beta}}_{NT}(\hat{\lambda}^{SL\,(b)})$ and $\tilde{\boldsymbol{\beta}}_{NT}(\hat{\lambda}^{C\,(b)})$.

Step 4. Repeat Steps 1-3 B times.

Step 5. Take the empirical quantiles from the resulting bootstrap distribution.

Bootstrap Validity

The bootstrap procedure described above is nonparametric, since it simply resamples data rows and estimating weighted estimator B times. To see whether the bootstrap distribution indeed gives a good estimate of the estimator distribution, we perform a Monte Carlo study based on the data. In particular, we generate data with the same characteristics as in the original data. Since correlation between the original regressors is negligible, we draw independent categorical variables with the same class distribution as in the original data for all but the property value, which we draw from a normal with the same mean and variance. Then we construct the total net sales data points by multiplying each with the OLS estimates obtained before. The observational OLS estimator is biased, but does provide the best fit, meaning that it is appropriate for generating similar outcomes. More specifically, for the experimental dataset we construct simulated values of total net sales $\tilde{s}_i$ by

$$\tilde{s}_i = \tilde{\boldsymbol{X}}_i \hat{\boldsymbol{\beta}}_N + \tilde{u}_i \qquad i = 1, \ldots, N, \tag{18}$$

where $\tilde{u}_i \sim N(0, \hat{\sigma}_u^2)$ and $N = 110$ like in the original data. The observational data is drawn analogously. Using this simulated data set, we estimate the COSE for both SL and C weights and compare the distributions with those of the bootstrapped values in the application.

With $M = B = 10\,000$ (bootstrap) simulations, we confirm a similarity between the two distributions with a Kolmogorov–Smirnov test statistic of 0.03 and a p-value of 0.000 for both weights. The

similarity is also visible in Fig. 7 and in the quantiles in Table 5. Most importantly, the quantiles that determine the significance of the results in the application (5% for SL, 10% for C), are also negative in the simulated distribution.

Figure 7: Bootstrap vs. simulated distributions of COSE with SL and C weights. These bootstrap histograms are for the weighted estimates of `discount_given` ($\tilde{\beta}_{TN,1}(\hat{\lambda})$) in particular.

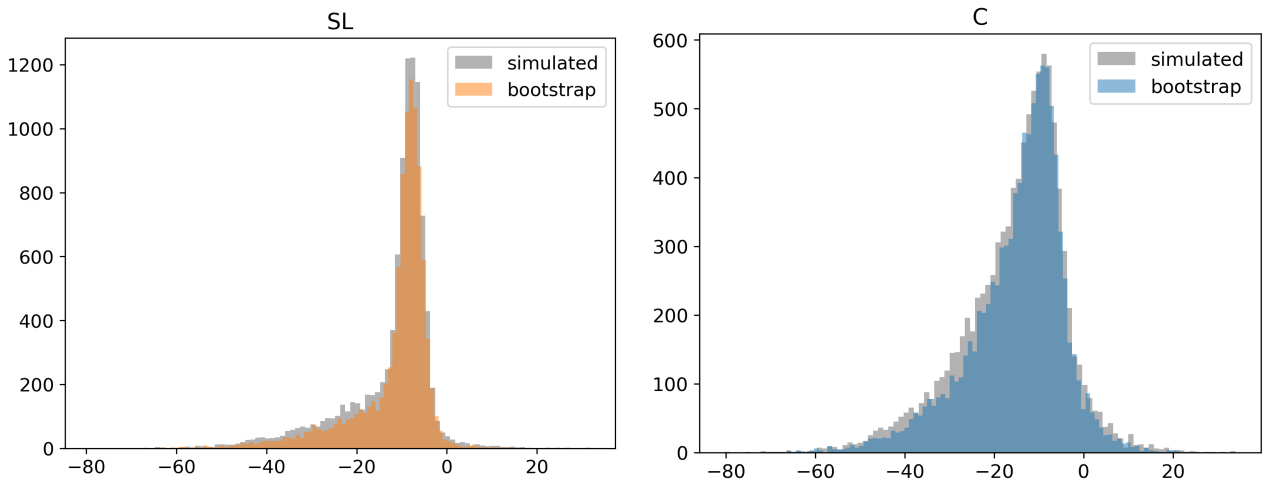


Table 5: Similarity of bootstrap and simulated distributions: quantiles and Kolmogorov-Smirnov test results. Relevant quantiles in application denoted in bold.

		0.01	0.05	0.1	0.90	0.95	0.99	KS stat.	KS p-val
SL	boot	-45.02	-30.99	-23.63	-4.97	**-3.81**	0.34	0.0352	0.000***
	sim	-46.76	-33.73	-26.08	-4.93	**-3.57**	4.01		
C	boot	-48.17	-35.90	-29.48	**-4.31**	-1.07	6.94	0.0331	0.000***
	sim	-49.71	-38.03	-31.28	**-3.64**	0.14	10.06		