Master Thesis Business Analytics and Quantitative Marketing

The Robustness of Causal Econometric and Machine Learning Methods

Author:	Mats Odekerken (455343)
Supervisor:	Mikhail Zhelonkin
Second Assessor:	Andrea Naghi

December 26, 2020

Abstract

Standard treatment evaluation methods typically rely on ordinary least squares, which is sensitive to outliers. Outliers and missing values are likely to be present in the datasets used, but the consequences are scarcely researched. Lately, such methods are also being designed in the field of machine learning, but also without emphasis on outliers. In this paper we examine the performances of treatment evaluation methods when the data is contaminated with both outliers and missing values. In particular, we investigate the instrumental variables and difference-in-differences methods. For the machine learning methods, we examine the debiased machine learning and approximate residual balancing methods. We propose robust alternatives for the instrumental variables and difference-in-differences methods as they rely on ordinary least squares. We see that the robust difference-in-differences method is superior when the data is contaminated. The robust instrumental variables method only outperforms the instrumental variables method when the data closely resembles to an elliptical structure. We apply the difference-in-differences type of methods to the data used in Card and Krueger (1993), and conclude that vertical outliers are present. The instrumental variables type of methods are applied to the data used in Dinkelman (2011). Although we detect outliers, no proposed alternative outperforms the instrumental variables method.

Keywords: Outliers, missing values, imputation, robust, treatment

Contents

1 Introduction

Causal inference is an important research area in the field of econometrics. In general, it is of importance to differentiate between causation and correlation. Within the field of causal inference it is of interest to determine whether treatment causes a change in the response, and to what extent. Determining these causal effects is however a task which is easier said than done. The fundamental problem of causal inference states that it is impossible to measure the treatment effect for a single observation, as one potential outcome is always unobserved (Rubin, 1974; Holland, 1986). To overcome this problem, numerous methods have been designed in the past with some being popular and often used to this day.

As data is nowadays increasingly gathered, more problems arise when it comes to the quality of the data obtained. In general, a sample with a large number of observations is preferred over a sample with a small number of observations, but that is under the assumption of well-behaved data. With an increase in the number of observations and/or dimensions, outliers and missing values are phenomena which become increasingly likely to appear. These are problems which occurred less often back in the days, and were therefore paid less attention to. Standard methods in all fields, but also within causal inference, were hence designed based on data assumed to behave as expected.

In our research, we focus on two of the most popular treatment evaluation methods, namely the Instrumental Variables (IV) and Difference-in-Differences (DiD) methods (Cameron and Trivedi (2005), Chapter 25). Estimation within both of these methods is traditionally based on Ordinary Least Squares (OLS) regression, a method which is proven to be non-resistant against outliers (Rousseeuw and Leroy, 1987). It has a breakdown point of 0%, meaning that one outlier can cause OLS to break down, leading to deceptive results (Donoho and Huber, 1983). Modern datasets are often high-dimensional and/or contain a large amount of observations, which almost guarantees the presence of at least one outlier. High-dimensional means $n < p$ in this case, where n and p stand for the number of observations and regressors respectively. In such cases, making use of OLS is not straightforward. Causal effect estimates may be biased, leading to erroneous policies for example. Hence, dealing with outliers is of great importance, making the use of robust regression techniques more attractive.

The main topic of our research is therefore to investigate the effect of outliers and missing values on treatment evaluation methods, a topic which is scarcely researched. Due to the well-researched robustness properties of OLS, treatment evaluation methods break down when outliers are present. We are therefore also interested in examining robust alternatives and their performances. For the IV method, a robust alternative is proposed in Freue et al. (2013), which is a method resistant to outliers due to the use of robust covariance matrices instead of sample covariance matrices. We extend the literature by investigating the performance of this method within a causal inference setting, which we call the Robust Instrumental Variables (RIV)

method. The extension specifically comes down to applying the method proposed in Freue et al. (2013) to a problem with a binary endogenous variable. We also investigate a different robust alternative in this paper, namely IV combined with the DetectDeviatingCells (DDC) algorithm (Rousseeuw and Bossche, 2018), which we will from now on refer to as the IV-DDC method. The RIV method is however the only method of which the robustification is purely model-based.

Designing a robust alternative for the DiD method is something which is done in Han et al. (2018), the focus of their research however differs from ours. They mainly focus on vertical outliers, while we are additionally interested in the effect of bad leverage points. Parallel trends between the treatment and control groups is usually the only criterion which is checked for before applying the DiD method. Outliers can be present in the controls, with the parallel trend assumption still being satisfied. Besides, vertical outliers can be witnessed after treatment. We propose the Robust Difference-in-Differences (RDiD) method by combining the general framework used in the DiD method with a robust regression technique.

Causal inference is a topic which has already been popular in econometrics for some time, it is however a relatively new research area in Machine Learning (ML). While there is not an incredible amount of interest yet, it is gaining the attention of researchers in ML at a high pace. Because the development of ML methods which suit causal inference type of problems is in an early stage, checking these methods' robustness properties with respect to outliers is something which is currently not focused on. This is where we want to step in and extend the literature, by evaluating the performances of these methods when outliers are present. Results from our paper may inspire researchers in ML, as robust ML methods for causal inference are not yet deliberately being designed.

We investigate robustness properties of two of the most popular ML methods designed for causal inference, namely the Debiased Machine Learning (DML) and Approximate Residual Balancing (ARB) methods, respectively proposed in Chernozhukov et al. (2018) and Athey et al. (2018). Within the DML method, there are two models which we examine, namely the Partial Linear Regression (PLR) and the Partial Linear Instrumental Variables (PLIV) models. These models are specifically designed for obtaining Average Treatment Effects (ATEs). There are methods available for obtaining Heterogeneous Treatment Effects (HTEs), but we leave the investigation of the robustness properties of these methods for further research. All econometric methods are designed for calculating ATEs, the current choice of ML methods makes it easier to make comparisons across methods and draw general conclusions based on the findings.

This paper is structured as follows: In Section 2 we discuss all literature which is relevant for the methods which we examine. In Section 3 we define all methods which we explore in our research. Section 4 contains details about the simulation study which we conduct to study the performances of all methods when the data is contaminated. In Section 5 we give a detailed description of the real datasets which we use to demonstrate the application of some methods in

practice. The results obtained from the simulation study and the real data are given in Section 6. Finally, we discuss our findings in Section 7. Additionally, we mention the limitations of our research in this section, as well as interesting topics for further research.

2 Literature

In this section all literature which is relevant for our research is described. The DiD method is popular in practice and has already been around for a while, but papers dedicated to exploring this method and its details have been around for a shorter period of time. One of the first papers where the approach followed was actually recognised to be a DiD method is the study conducted by Card and Krueger (1993). By controlling for omitted variables they come to a conclusion differing from what would have been concluded based on traditional economic theory. They illustrate employment growth due to higher minimum wages, while theory suggests a decrease in the employment level in such a case. The DiD method was studied more extensively after this paper, its characteristics have been thoroughly investigated since.

A pitfall of the DiD method which arised later on is that standard errors are underestimated in some cases. In Bertrand et al. (2004) this problem is proven to be present in case of a serially correlated outcome measured over multiple time periods. Donald and Lang (2007) in turn demonstrate the presence of this problem if the number of groups is small. In practice, this problem is typically dealt with by clustering the standard errors. However, in the case of a small amount of groups, Donald and Lang (2007) show that this solution may yield even more inaccurate inference. When examining the real data, we overcome underestimation by combining clustering with the bootstrap as in Cameron et al. (2008).

The IV method was first proposed in Wright (1928), strangely enough in an appendix. This method is particularly useful in the field of causal inference, as it deals with the common phenomenon of endogenous variables (Cameron and Trivedi (2005), Chapter 4.8). In such cases, OLS estimation results in biased parameter estimates, leading to a biased ATE. IV estimation has been shown to successfully tackle the problem of endogenous variables in a causal inference setting in Angrist et al. (1996). It is important for an instrument to have a considerable correlation with the endogenous variable, the IV parameter estimates will otherwise be biased towards the OLS estimates (Bound et al., 1995).

The problem of both the standard DiD and IV methods is that they are not resistant to outliers, as they typically make use of OLS. Hence, robust alternatives of these methods are preferred for contaminated datasets. As mentioned in Section 1, we propose the RDiD method as a robust alternative to the DiD method. The regression technique which it is based on is MM estimation (Yohai, 1987), a method which is both robust and efficient. The method consists of a combination of the S-estimator (Huber, 1992) and the M-estimator (Huber et al., 1973), these methods are known for their robustness and efficiency respectively. The RIV method which we examine

comes down to a natural robustification of the IV method. The IV estimator can be decomposed into multiple sample covariance matrices. As the sample estimator does not robustly estimate the covariance, the idea of Freue et al. (2013) is to obtain parameter estimates by making use of robustly estimated covariance matrices. We make use of the Minimum Covariance Determinant (MCD) estimator in order to estimate the covariance matrices (Rousseeuw, 1985).

More and more researchers in ML are starting to see the relevance and importance of causality. Chernozhukov et al. (2018) propose the DML method which is partially based on the econometric semi-parametric method proposed by Robinson (1988). This method makes use of regularization, making it suitable for high-dimensional problems. As the standard DiD and IV methods can not be applied instantly for high-dimensional problems, ML type of methods are more attractive in such cases. The main idea behind the DML method is to tackle regularization bias by making use of a doubly robust approach.

The DML method requires the propensity score to be consistently estimable for the ATE to be $\sqrt{n}$ consistently estimated, where n stands for the amount of observations. Another popular ML method which does not impose this restriction is the ARB method (Athey et al., 2018). This method actually resembles to the DML method, as it also makes use of regularization. Also, both methods build upon the work of Robinson (1988). The ARB method is however less restrictive as already mentioned, and therefore more widely applicable.

The clear advantage of the ML methods is that they can deal with high-dimensional problems. Besides, researchers often select a fraction of the covariates based on common sense and econometric intuition, possibly leaving out relevant variables. If there is uncertainty regarding the relevance of some of the explanatory variables, choosing one of the ML methods is desirable. The parameters of the covariates are shrunk towards zero due to regularization if they do not affect the outcome.

3 Methodology

In this section all methods used for our research are explained in detail. We differentiate between ML and econometric methods by explaining both types of methods in separate sections. The econometric methods are explained in Section 3.1, with the proposed robust alternatives being explained in Section 3.2. The ML methods are explained in Section 3.3. We let n denote the amount of individuals throughout the rest of the paper, subscript i holds for $i = 1, \ldots, n$ if no additional comments are made.

3.1 Standard Econometric Methods

In this section the standard econometric methods used for our research are explained in detail. The DiD method is discussed in Section 3.1.1, the IV method in Section 3.1.2.

3.1.1 Difference-in-Differences

The DiD method determines treatment effects by observing a group of individuals over different periods of time. The simplest case, which we use for our research, is observing individuals over two time periods. The main idea of the DiD estimator is to determine the treatment effect by comparing the average change of the outcome between the treatment and control groups over time. The DiD method is illustrated in Figure 1, where $t = 0$ and $t = 1$ stand for observations measured before and after treatment respectively.

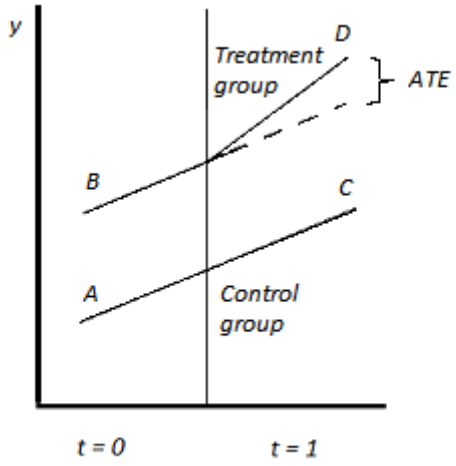

Figure 1: Illustration of the DiD method.

Figure 1 shows the average treatment effect given as $ATE = \big(y_D - y_B\big) - \big(y_C - y_A\big)$, it can be interpreted as the difference in outcome with respect to the outcome expected based on the trends. The control group is not treated, its trend is therefore assumed to remain stable after treatment of the treatment group. Figure 1 shows parallel trends, the ATE however does not closely resemble the actual treatment when these trends are not parallel. In order for the DiD method to yield outcomes which make sense, the assumption of parallel trends must be satisfied. If this assumption is violated, the ATE could stem solely from differing trends.

For the DiD method we define y_i as the outcome of individual i, t_i indicates whether the time period is pre or post treatment by taking on values zero and one respectively. Vector $\boldsymbol{x}_i = \big(x_{i1}, \ldots, x_{ik}\big)'$ contains the controls, d_i indicates whether an individual is treated or not given by values zero and one respectively. The final model is defined as

$$y_i = \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta} + \beta_{k+1}d_i + \beta_{k+2}t_i + \beta_{k+3}(d_i \cdot t_i) + \epsilon_i, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ and ϵ_i is an unobserved error term. The main parameter of interest is β_{k+3}, representing the treatment effect. The name of the DiD method stems from the way this

parameter is estimated, a derivation is given as

$$\begin{aligned}\widehat{\beta}_{k+3} &= \big(\mathbb{E}[y_i|d_i=1,t_i=1]-\mathbb{E}[y_i|d_i=1,t_i=0]\big)-\big(\mathbb{E}[y_i|d_i=0,t_i=1]-\mathbb{E}[y_i|d_i=0,t_i=0]\big)\\ &= \big((\beta_0+\boldsymbol{x}_i'\boldsymbol{\beta}+\beta_{k+1}+\beta_{k+2}+\beta_{k+3}\big)-\big(\beta_0+\boldsymbol{x}_i'\boldsymbol{\beta}+\beta_{k+1}\big)\big)-\big((\beta_0+\boldsymbol{x}_i'\boldsymbol{\beta}+\beta_{k+2})-(\beta_0+\boldsymbol{x}_i'\boldsymbol{\beta})\big)\\ &= \big(\beta_{k+2}+\beta_{k+3}\big)-\beta_{k+2} \qquad (2)\\ &= \beta_{k+3}.\end{aligned}$$

Clearly, Equation (2) shows that the estimate of the treatment effect is obtained by calculating a difference of differences.

3.1.2 Instrumental Variables

The IV method is widely used and also useful for obtaining causal estimates. A common problem within econometrics is obtaining biased parameters due to omitted variables, see for example Chapter 4.7 of Cameron and Trivedi (2005). IV estimation tackles this problem by using instruments, an instrument is valid if it is correlated with the endogenous explanatory variable and uncorrelated with the error term. The latter condition implies that the instrument should be uncorrelated with any omitted variable, as these are captured within the error term.

For IV estimation we make use of only one time period, hence we adapt the model as in Equation (1) slightly. The model used for IV estimation is defined as

$$y_i = \alpha_0 + \boldsymbol{x}_i'\boldsymbol{\alpha} + \alpha_{k+1}d_i + \nu_i, \qquad (3)$$

where $\boldsymbol{\alpha} = \big(\alpha_1,\ldots,\alpha_k\big)'$ and ν_i is an unobserved error term. It is common for the treatment variable d_i to be endogenous, that is $\rho_{d_i\nu_i} \neq 0$, meaning that we need to find a valid instrument for this variable. Instrument z_i is valid if $\rho_{z_i\nu_i} = 0$ and $\rho_{z_i d_i} \neq 0$. IV estimation is also known as Two-Stage Least Squares $\big($TSLS$\big)$, as the final parameter estimates are obtained by sequentially running two regressions. Due to the large amount of Greek letters needed for definitions in this paper, we describe the TSLS procedure without defining additional models. In the first stage, the endogenous variable is regressed on all exogenous variables and the instrument. From this regression, fitted values for the endogenous variable are obtained. In the second stage, the outcome is regressed on all exogenous variables and the fitted values obtained from the first stage.

3.2 Robust Econometric Methods

As our research deals with investigating robustness properties, we are interested in the differences in performance of a method and its proposed robust alternative. Ideally, robust methods prove to perform better in settings with outliers, indicating that using this method's non-robust version in such cases leads to less credible results. In this section we propose alternatives which we expect to be more robust compared to the standard econometric methods. The RDiD method is proposed in Section 3.2.1, the RIV and IV-DDC methods are proposed in Section 3.2.2.

3.2.1 Robust Difference-in-Differences

The RDiD method which we propose comes down to applying a robust estimator to the model as defined in Equation (1). We make use of the MM-estimator as already mentioned, it is one of the most popular robust regression techniques in practice. Using such a robust estimator is in essence a safe option, if the percentage of outliers does not attain the breakdown point. Always using a robust alternative is however not an approach which is favoured. It is well-known that OLS is efficient if all of its assumptions are met, meaning that robust regression is not optimal in such cases.

If we continue on Equation (1) where the general DiD model is specified, we can define the corresponding MM-estimator as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\mathrm{MM}} &= \underset{b_0,\ldots,b_{k+3}}{\arg\min} \sum_{i=1}^{n} \rho_2\left(\frac{y_i - b_0 - \boldsymbol{x}_i'\boldsymbol{b} - b_{k+1}d_i - b_{k+2}t_i - b_{k+3}\big(d_i \cdot t_i\big)}{\widehat{\sigma}_{\mathrm{S}}}\right) \\ &= \underset{\boldsymbol{b}^*}{\arg\min} \sum_{i=1}^{n} \rho_2\left(\frac{y_i - \boldsymbol{x}_i^{*\prime}\boldsymbol{b}^*}{\widehat{\sigma}_{\mathrm{S}}}\right),\end{aligned} \tag{4}$$

where $\boldsymbol{x}_i^* = \big(1, x_{i1}, \ldots, x_{ik}, d_i, t_i, (d_i \cdot t_i)\big)'$ and $\boldsymbol{b}^* = \big(b_0, \ldots, b_{k+3}\big)'$. Preliminary scale estimate $\widehat{\sigma}_{\mathrm{S}} = \widehat{\sigma}_{\mathrm{M}}\big(\widehat{\boldsymbol{\beta}}_{\mathrm{S}}\big)$ is obtained from

$$\widehat{\boldsymbol{\beta}}_{\mathrm{S}} = \underset{\boldsymbol{b}^*}{\arg\min}\ \widehat{\sigma}^2_{\mathrm{M}}(\boldsymbol{b}^*). \tag{5}$$

Estimating $\widehat{\sigma}_{\mathrm{M}}\big(\boldsymbol{b}^*\big)$ in turn solves

$$\frac{1}{n}\sum_{i=1}^{n} \rho_1\left(\frac{y_i - \boldsymbol{x}_i^{*\prime}\boldsymbol{b}^*}{\widehat{\sigma}_{\mathrm{M}}\big(\boldsymbol{b}^*\big)}\right) = \delta, \tag{6}$$

where ρ_1 represents a loss function, $\delta = \mathbb{E}\Big[\rho_1\big(Z\big)\Big]$ with $Z \sim N\big(0,1\big)$.

The loss function is denoted by $\rho_2(\cdot)$ in Equation (4), we make use of the Tukey bisquare function for both ρ_1 and ρ_2, that is

$$\rho_i\big(x\big) = \begin{cases} \dfrac{x^6}{6c_i^4} - \dfrac{x^4}{2c_i^2} + \dfrac{x^2}{2} & \text{if } \big|x\big| \leq c_i, \text{ for } i = 1,2, \\ \dfrac{c_i^2}{6} & \text{if } \big|x\big| > c_i, \text{ for } i = 1,2, \end{cases} \tag{7}$$

where c_i is a tuning constant. We set $c_1 = 1.548$ and $c_2 = 4.685$, yielding a breakdown point of 50% for $\widehat{\boldsymbol{\beta}}_{\mathrm{S}}$ and 95% efficiency for $\widehat{\boldsymbol{\beta}}_{\mathrm{MM}}$. Analytically solving Equation (4) leads to

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \boldsymbol{x}_i^{*\prime}\boldsymbol{b}^*}{\widehat{\sigma}_{\mathrm{S}}}\right)\boldsymbol{x}_i^* = 0, \tag{8}$$

where

$$\begin{aligned}\psi(x) &= \rho_2'(x) \\ &= \begin{cases} \dfrac{x^5}{c_2^4} - \dfrac{2x^3}{c_2^2} + x & \text{if } \big|x\big| \leq c_2, \\ 0 & \text{if } \big|x\big| > c_2. \end{cases}\end{aligned} \tag{9}$$

The condition in Equation (8) can be rewritten, $\widehat{\boldsymbol{\beta}}_{\mathrm{MM}}$ is the solution of

$$\sum_{i=1}^{n} w_i\Big(y_i - \boldsymbol{x}_i^{*'}\boldsymbol{b}^*\Big)\boldsymbol{x}_i^* = 0, \tag{10}$$

where

$$w_i = w\big((y_i - \boldsymbol{x}_i^{*'}\boldsymbol{b}^*)/\widehat{\sigma}_{\mathrm{S}}\big) = \frac{\psi\big((y_i - \boldsymbol{x}_i^{*'}\boldsymbol{b}^*)/\widehat{\sigma}_{\mathrm{S}}\big)}{\big(y_i - \boldsymbol{x}_i^{*'}\boldsymbol{b}^*\big)/\widehat{\sigma}_{\mathrm{S}}}. \tag{11}$$

This procedure can be seen as a problem which can be solved by applying Iteratively Reweighted Least Squares $\big(\mathrm{IRLS}\big)$, with w_i as the weights. If $\widehat{\boldsymbol{\beta}}_t$ is given during step t of this procedure, updates are executed as follows: the weights are updated according to Equation (11), that is $w_i = w\big((y_i - \boldsymbol{x}_i^{*'}\widehat{\boldsymbol{\beta}}_t)/\widehat{\sigma}_{\mathrm{S}}\big)$. Subsequently, the parameter estimates are updated according to

$$\widehat{\boldsymbol{\beta}}_{t+1} = \arg\min_{\boldsymbol{b}^*} \frac{1}{n}\sum_{i=1}^{n} w_i\big(y_i - \boldsymbol{x}_i^{*'}\boldsymbol{b}^*\big)^2. \tag{12}$$

3.2.2 Robust Instrumental Variables

The robust variants of the IV method which we investigate are explained in this Section. The RIV method relies on an estimation procedure from which the framework is equal to that of the IV method.

Consider the regression model as in Equation (3) given in matrix form, that is

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_0 \end{bmatrix} + \begin{bmatrix} x_{11} & \dots & x_{1k} & d_1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{n1} & \dots & x_{nk} & d_n \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_{k+1} \end{bmatrix} + \begin{bmatrix} \nu_1 \\ \vdots \\ \nu_n \end{bmatrix}, \tag{13}$$

which can be compactly written as $\boldsymbol{y} = \boldsymbol{\alpha}_0 + \boldsymbol{X}\boldsymbol{\alpha}^* + \boldsymbol{\nu}$. The OLS estimates of the parameters are given as

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}^*_{\mathrm{OLS}} &= \big(\boldsymbol{X}'\boldsymbol{X}\big)^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= \big(n\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}\boldsymbol{X}}\big)^{-1} n\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}\boldsymbol{y}},\end{aligned} \tag{14}$$

$$\widehat{\boldsymbol{\alpha}}_{0,\mathrm{OLS}} = \mu_{\boldsymbol{y}} - \boldsymbol{\mu}'_{\boldsymbol{X}}\widehat{\boldsymbol{\alpha}}^*_{\mathrm{OLS}}, \tag{15}$$

where $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{A}\boldsymbol{B}}$ stands for the estimated covariance between $\boldsymbol{A}$ and $\boldsymbol{B}$, $\mu_{\boldsymbol{y}}$ and $\boldsymbol{\mu}_{\boldsymbol{X}}$ stand for the estimated location of $\boldsymbol{y}$ and $\boldsymbol{X}$ respectively. When one or more of the explanatory variables

are endogenous, parameter estimates obtained by OLS become biased. If there are valid instruments available, IV is a useful alternative method. Endogeneity of d_i is the central problem in causal inference, we build on that principle throughout the rest of this section.

In our simulation study, as well as with analyzing the real data, we use a single instrument z_i. Hence, we limit all further definitions to cases where d_i is solely instrumented by this variable. Freue et al. (2013) propose a way of estimating models which contain binary exogenous variables, we however assume all exogenous variables to be continuous in the rest of this section. As the real data which we examine does contain binary exogenous variables, we demonstrate the procedure proposed in Freue et al. (2013) in Section 6.2 by directly applying it to a real dataset.

The matrix containing the instrument is constructed as

$$\boldsymbol{Z} = \begin{bmatrix} x_{11} & \dots & x_{1k} & z_1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{n1} & \dots & x_{nk} & z_n \end{bmatrix}, \tag{16}$$

all exogenous explanatory variables serve as their own instruments in $\boldsymbol{Z}$. We define the analogous estimation procedure as described in Equation (3) for the model given in matrix notation. For the first stage we define

$$\widetilde{\boldsymbol{x}}_j = \boldsymbol{Z}\boldsymbol{\pi} + \boldsymbol{\eta}, \qquad j = 1, \dots, k+1, \tag{17}$$

where $\widetilde{\boldsymbol{x}}_j = \big(x_{1j}, \dots, x_{nj}\big)'$, $\boldsymbol{\pi} = \big(\pi_1, \dots, \pi_{k+1}\big)'$ and $\boldsymbol{\eta} = \big(\eta_1, \dots, \eta_n\big)'$. For convenience, d_i is contained in $\widetilde{\boldsymbol{x}}_j$ as x_{ik+1}.

We apply OLS to the model in Equation (17) for all covariates, the matrix of fitted values of all these regressions is given as

$$\begin{aligned} \hat{\boldsymbol{X}} &= \boldsymbol{Z}\widehat{\boldsymbol{\pi}}_{\text{OLS}} \\ &= \boldsymbol{Z}\big(\boldsymbol{Z}'\boldsymbol{Z}\big)^{-1}\boldsymbol{Z}'\boldsymbol{X} \\ &= \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}. \end{aligned} \tag{18}$$

It holds that $\boldsymbol{P}_{\boldsymbol{Z}}^2 = \boldsymbol{P}_{\boldsymbol{Z}}$ and $\boldsymbol{P}_{\boldsymbol{Z}}' = \boldsymbol{P}_{\boldsymbol{Z}}$, respectively meaning that $\boldsymbol{P}_{\boldsymbol{Z}}$ is both idempotent and symmetric. Derivations of these results can be found in Equations 72 and 73 in Section A. For the second stage we make use of

$$\boldsymbol{y} = \boldsymbol{\Delta}_0 + \hat{\boldsymbol{X}}\boldsymbol{\Delta} + \boldsymbol{\Lambda}, \tag{19}$$

where $\boldsymbol{\Delta} = \big(\Delta_1, \dots, \Delta_{k+1}\big)'$, $\boldsymbol{\Lambda} = \big(\Lambda_1, \dots, \Lambda_n\big)'$ and $\boldsymbol{\Delta}_0 = \big(\Delta_0, \dots, \Delta_0\big)'$, that is $\boldsymbol{\Delta}_0 \in \mathbb{R}^n$.

The IV estimators are obtained by applying OLS to Equation (19), that is

$$\begin{aligned}\widehat{\boldsymbol{\Delta}}_{\text{IV}} &= \left(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}}\right)^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y} \\ &= \left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{XZ}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{ZZ}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{ZX}}\right)^{-1}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{XZ}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{ZZ}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Zy}},\end{aligned} \tag{20}$$

$$\widehat{\boldsymbol{\Delta}}_{0,\text{IV}} = \mu_{\boldsymbol{y}} - \boldsymbol{\mu}'_{\hat{\boldsymbol{X}}}\widehat{\boldsymbol{\Delta}}_{\text{IV}}. \tag{21}$$

A full derivation of Equation (20) is given in Equation (74) in Section A. Such covariance matrices are traditionally estimated by calculating sample covariances, a method which is not resistant against outliers. Freue et al. (2013) propose to simply replace these sample covariance matrices by robust estimates of the covariance matrix, they for example apply the S-estimator (Rousseeuw and Yohai, 1984).

In our research, we make use of a more popular robust estimator, the MCD estimator to be precise. The main idea of the estimator is to search for the $h \leq n$ points resulting in a minimal value of the determinant of the covariance matrix. Estimators of location and scatter are calculated based on these h points, filtering out all outliers in the ideal case. The subset size is calculated as $h = \pi n$, where $0.5 \leq \pi \leq 1$. The MCD estimator achieves maximum robustness when $\pi = 0.5$, we however set $\pi = 0.75$ in order to avoid multicollinearity issues.

We define the estimator of location of subset H obtained from the variables of interest as

$$\widehat{\boldsymbol{\mu}}_H = \frac{1}{h}\sum_{i\in H}\check{\boldsymbol{x}}_i. \tag{22}$$

where $\check{\boldsymbol{x}}_i = \left(y_i, x_{i1}, \ldots, x_{ik}, d_i, z_i\right)'$. The estimator of scatter is in turn defined as

$$\widehat{\boldsymbol{\Sigma}}_H = \frac{1}{h}\sum_{i\in H}\left(\check{\boldsymbol{x}}_i - \widehat{\boldsymbol{\mu}}_H\right)\left(\check{\boldsymbol{x}}_i - \widehat{\boldsymbol{\mu}}_H\right)'. \tag{23}$$

All covariance matrices used in Equation (20) are extracted from $\widehat{\boldsymbol{\Sigma}}_H$. Subset H is in term determined by solving

$$H = \underset{\widetilde{H}:\left|\widetilde{H}\right|=h}{\arg\min}\det\left(\widehat{\boldsymbol{\Sigma}}_{\widetilde{H}}\right). \tag{24}$$

The estimator of location is fisher consistent under a normal distribution, the estimator of scatter however has to be corrected in order to achieve fisher consistency. Note that the assumption of a normal distribution is not appropriate as d_i is binary, but we elaborate on this issue later on in this section.

The raw estimates of location and scatter are given as $\widehat{\boldsymbol{\mu}}_{\text{raw}} = \widehat{\boldsymbol{\mu}}_H$ and $\widehat{\boldsymbol{\Sigma}}_{\text{raw}} = c_x c_{nk+2}\widehat{\boldsymbol{\Sigma}}_H$ respectively, where c_{nk+2} is as in Pison et al. (2002) and

$$c_x = \frac{x}{F_{\Gamma\left(\frac{(k+2)}{2}+1,1\right)}\left(\frac{\chi_x^2}{2}\right)}. \tag{25}$$

As h has to be manually chosen, it is likely to differ from the actual amount of good data points. If h is initialized too high, the MCD estimator will make use of a subset still including outliers. However, if h is chosen too small, the quality of the location and scatter estimates is not maximal, as some useful data points are left out. In order to gain efficiency, a reweighting step is applied based on the initial MCD estimator. This reweighting step is based on the Mahalanobis Distance $\big(\text{MD}\big)$, which is defined as

$$MD\big(\check{\boldsymbol{x}}_i, \widehat{\boldsymbol{\mu}}_{\text{raw}}, \widehat{\boldsymbol{\Sigma}}_{\text{raw}}\big) = \sqrt{\big(\check{\boldsymbol{x}}_i - \widehat{\boldsymbol{\mu}}_{\text{raw}}\big)' \widehat{\boldsymbol{\Sigma}}_{\text{raw}}^{-1} \big(\check{\boldsymbol{x}}_i - \widehat{\boldsymbol{\mu}}_{\text{raw}}\big)} \tag{26}$$

for our model.

It holds that $MD^2\big(\check{\boldsymbol{x}}_i, \widehat{\boldsymbol{\mu}}_{\text{raw}}, \widehat{\boldsymbol{\Sigma}}_{\text{raw}}\big) \sim \chi^2(k+3)$ if $\check{\boldsymbol{x}}_i$ comes from a normal distribution with $\widehat{\boldsymbol{\mu}}_{\text{raw}}$ and $\widehat{\boldsymbol{\Sigma}}_{\text{raw}}$ as location and scatter respectively. If an observation has a large MD, its location and scatter lie far from the population's locations and scatter, meaning that it is likely to concern an outlier. The observations are weighted based on this relationship, that is

$$w_i = \begin{cases} 1 & \text{if } MD^2\big(\check{\boldsymbol{x}}_i, \widehat{\boldsymbol{\mu}}_{\text{raw}}, \widehat{\boldsymbol{\Sigma}}_{\text{raw}}\big) \leq \chi^2_{1-\delta}(k+3), \\ 0 & \text{if } MD^2\big(\check{\boldsymbol{x}}_i, \widehat{\boldsymbol{\mu}}_{\text{raw}}, \widehat{\boldsymbol{\Sigma}}_{\text{raw}}\big) > \chi^2_{1-\delta}(k+3), \end{cases} \tag{27}$$

where $\chi^2_{1-\delta}(\cdot)$ stands for the $\big(1-\delta\big)$ quantile of the corresponding distribution, we set $\delta = 0.025$. The reweighted estimates of location and scatter are now given as

$$\widehat{\boldsymbol{\mu}}_{\text{weight}} = \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i \check{\boldsymbol{x}}_i \tag{28}$$

and

$$\widehat{\boldsymbol{\Sigma}}_{\text{weight}} = \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i \big(\check{\boldsymbol{x}}_i - \widehat{\boldsymbol{\mu}}_{\text{weight}}\big)\big(\check{\boldsymbol{x}}_i - \widehat{\boldsymbol{\mu}}_{\text{weight}}\big)' \tag{29}$$

respectively. Estimation is carried out by making use of the FAST-MCD method (Rousseeuw and Driessen, 1999).

An important thing to note about robustly estimating a covariance matrix in general, is that a data matrix containing one or more categorical variables is not ideal. For the MCD estimator which we employ specifically, it searches for a subset which results in the smallest determinant of the covariance matrix. This determinant is in turn proportional to the volume of an ellipsoid, meaning that a data matrix of which the points can be fit into an ellipsoid is desirable. Logically, categorical variables do not fit this picture. Besides, the estimates of location and scatter are only fisher consistent under a normal distribution, and the MDs also assume the data to be elliptically structured.

In Freue et al. (2013) an alternative estimation procedure is proposed if the data consists of both continuous and binary variables, this however does not necessarily solve our problem. This particular method can deal with exogenous dummy covariates, within causal inference it is however common for the binary treatment variable to be endogenous. Extending the procedures as in Freue et al. (2013) for endogenous binary variables is however not possible, as the main formula given in Equation (20) relies upon correlations between the endogenous variables and the instruments. As we however think that the RIV method has the potential to be the best robust alternative, we are interested in investigating the method, even when the data is not a perfect fit for the method.

As already mentioned, the procedure proposed in Freue et al. (2013) for data matrices containing continuous and binary variables does not necessarily solve the endogeneity problem faced within causal inference, it might however be useful in some cases. A less common problem is characterized by an exogenous treatment variable, and one or more endogenous control variables, see Frölich (2008) for the discussion of some examples. If endogeneity is neglected in such a case, regression estimates will also lead to a biased estimate of the treatment effect. In Freue et al. (2013), the RIV method is proven to perform well for such data, meaning that it will be the go to robust alternative. As this is however a problem rarely encountered in practice, we do not include this scenario in our simulation study.

If the treatment variable is endogenous, the RIV method may not look like the most straightforward method to use. As the IV estimator given in Equation (20) can be estimated in a single step, replacing OLS regression by a robust technique may yield better estimates. Note however that $\hat{\boldsymbol{X}}$ consists of linear combinations of $\boldsymbol{Z}$ and $\boldsymbol{X}$, meaning that the presence of only a small fraction of outliers in these matrices is likely to result in more than half of the observations in $\hat{\boldsymbol{X}}$ to be outlying, meaning that robust regression also breaks down.

As TSLS equals IV estimation, a different idea is to sequentially apply robust regression in both of the stages. Applying robust linear regression in the first stage however fails, as the treatment variable is perfectly predicted. MM estimation omits all observations where $d_i = 0$ and sets all parameters to zero, except for the parameter of the constant which is set to one. Robust logistic regression techniques also exist (Rousseeuw and Christmann, 2003; Feng et al., 2014), applying such a technique to the first stage would however result in the forbidden regression as explained in Chapter 9.5.2 in Wooldridge (2010). Specifically, using robust logistic regression for the first stage would lead to fitted values $\widehat{d}_i$ coming from a nonlinear function. In this case, $\widehat{d}_i$ and the covariates are not guaranteed to be uncorrelated with the error terms of the second stage, while using OLS in the first stage does ensure this relation.

A different idea is to only robustify the second stage, while still using OLS in the first stage. Fitted values from the first stage are affected by bad leverage points in this case, but the second stage is robustly estimated. As mentioned in the preceding paragraph, a few bad leverage points can already cause this strategy to break down. As the strategies from both this and the preceding paragraphs are only expected to be robust in a limited amount of cases, we do not conduct further research on these methods.

Hence, there is only one strategy additional to the RIV method which we examine, which is the IV-DDC method. This method comes down to regularly applying the IV method, combined with an intermediate step taken to detect cellwise outliers. It is outside the scope of this paper to explain the DDC algorithm in detail, but the main idea of the method is to flag cells as outlying or not by making use of correlations between the variables. We apply the DDC algorithm before estimation, and set the cells detected as outlying to missing. Afterwards, we impute these missing values to come up with a complete data matrix. In Section 4.2 we give a more detailed description of the imputation strategy that goes along with our research.

3.3 Machine Learning Methods

In this section, the ML methods used for our research are explained in detail. Applying all previously defined methods becomes infeasible if the data is high-dimensional, that is $n < p$ where p stands for the amount of covariates. In such cases, regularization techniques can be applied for obtaining regression estimates. Both the DML and the ARB methods make use of regularization, meaning that they can be applied in a high-dimensional setting. As investigating cases where $n < p$ however becomes computationally too expensive, we do not include such cases in our simulation study. The DML method is discussed in Section 3.3.1, with the PLR and PLIV models corresponding to the DML method being explained in Sections 3.3.1.1 and 3.3.1.2 respectively. The ARB method is discussed in Section 3.3.2.

3.3.1 Debiased Machine Learning

The DML method proposed by Chernozhukov et al. (2018) overcomes problems which arise when ML techniques which rely on regularization are applied. Applying regularization possibly leads to bias in the estimators, which is logically related to the word Debiased in DML. Chernozhukov et al. (2018) show that almost all regularization bias can be overcome by making use of orthogonalization, which is a concept based on Neyman-orthogonal moment conditions (Neyman, 1959,9). Bias due to remaining terms is removed by making use of cross-fitting, which is a more efficient way of sample splitting. Orthogonalization is too difficult to explain in a non-technical way, we can however briefly describe the cross-fitting procedure before we dive into all theory of the DML method.

Cross-fitting is based on sample splitting, but it tackles the potential problem of efficiency loss which can arise when sample splitting is applied. The cross-fitting method also makes use of Neyman-orthogonal moments, but the main idea behind this method is to additionally swap certain samples. By applying this technique, we obtain multiple estimates of which the average is taken. The eventual outcome avoids a potential loss in efficiency, which is why it is preferred over the sample splitting method.

Our main interest lies in obtaining the true value of ω_0 of target parameter $\boldsymbol{\omega} \in \Omega$, where $\Omega \subset \mathbb{R}^{d_\omega}$ with d_ω denoting the dimension of $\boldsymbol{\omega}$. We assume that ω_0 satisfies moment conditions

$$\mathbb{E}\Big[\boldsymbol{\zeta}\big(W; \omega_0, \mu_0\big)\Big] = 0, \tag{30}$$

where W is a random variable and $\boldsymbol{\zeta} = \big(\zeta_1, \ldots, \zeta_{d_\omega}\big)'$ is a vector containing score functions. Parameter μ_0 denotes the true value of nuisance parameter $\mu \in T$, where T is a convex subset of a normed vector space. Neyman orthogonality is required for $\boldsymbol{\zeta}$, hence we introduce $\widetilde{T} = \big(\mu - \mu_0 : \mu \in T\big)$ and its pathwise derivative map $D_r : \widetilde{T} \to \mathbb{R}^{d_\omega}$ as

$$D_r\big[\mu - \mu_0\big] = \partial_r\bigg(\mathbb{E}\Big[\boldsymbol{\zeta}\big(W; \omega_0, \mu_0 + r(\mu - \mu_0)\big)\Big]\bigg), \quad r \in \big[0, 1\big). \tag{31}$$

For $r = 0$, the pathwise derivative map boils down to

$$D_0\big[\mu - \mu_0\big] = \partial_\mu \mathbb{E}\Big[\boldsymbol{\zeta}\big(W; \omega_0, \mu_0\big)\Big]\Big[\mu - \mu_0\Big]. \tag{32}$$

We furthermore define $\mathcal{T}_n \subset T$ as a nuisance realization set such that the probability of the estimators μ_0 taking on values in this set is high. Fulfilling the moment conditions in Equation (30) is one of the requirements for score function $\boldsymbol{\zeta}$ to meet the orthogonality condition at (ω_0, μ_0). If the pathwise derivative map as in Equation (31) additionally exists for $\mu \in \mathcal{T}_n$ and fades at $r = 0$, that is $D_0\big[\mu - \mu_0\big] = 0$, function $\boldsymbol{\zeta}$ is said to satisfy the orthogonality condition.

We continue our analysis with the application of cross-fitting to the data, a method which we explain in detail this time on. We first of all assume that we have a sample $\big\{W_i\big\}_{i=1}^n$ at our disposal, representing independent and identically distributed $\big($i.i.d.$\big)$ copies of random variable W. The procedure of cross-fitting is defined as follows:

First, take a random K-fold partition $\big\{I_k\big\}_{k=1}^K$ of individual indices $\big\{1, \ldots, n\big\}$. For each $k \in \big\{1, \ldots, K\big\}$, define $I_k^c = \big\{1, \ldots, n\big\} \setminus I_k$. For convenience, let $m = \frac{n}{K}$ denote the number of observations in each fold. An ML estimator of μ_0 is subsequently determined by calculating

$$\widehat{\mu}_{0,k} = \widehat{\mu}_0\bigg(\big\{W_i\big\}_{i \in I_k^c}\bigg), \tag{33}$$

where $\widehat{\mu}_{0,k}$ is a random element in T. The estimate of the target parameter is denoted by $\widetilde{\omega}_{0,k}$

and obtained by solving

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{m,k}\Big[\boldsymbol{\zeta}\big(W;\widetilde{\omega}_0,\widehat{\mu}_{0,k}\big)\Big]=0, \tag{34}$$

where $\mathbb{E}_{m,k}[\cdot]$ is the empirical expectation over data fold k, calculated as $\mathbb{E}_{m,k}\big[\boldsymbol{\zeta}(W)\big]=\frac{1}{m}\sum_{i\in I_k}\boldsymbol{\zeta}\big(W_i\big)$.

If Equation (34) can not be exactly solved, $\widetilde{\omega}_0$ is seen as an approximate solution if it satisfies

$$\left\|\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{m,k}\Big[\boldsymbol{\zeta}\big(W;\widetilde{\omega}_0,\widehat{\mu}_{0,k}\big)\Big]\right\|\leq\inf_{\omega\in\Omega}\left\|\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{m,k}\Big[\boldsymbol{\zeta}\big(W;\omega_0,\widehat{\mu}_{0,k}\big)\Big]\right\|+\widetilde{\epsilon}_m, \tag{35}$$

where $\widetilde{\epsilon}_m=o\big(\delta_m m^{-\frac{1}{2}}\big)$, with $\big(\delta_m\big)_{m\geq 1}$ representing some sequence of positive constants converging to zero. The final estimator is obtained by averaging over all partitions, that is

$$\widehat{\omega}_0=\frac{1}{K}\sum_{k=1}^{K}\widetilde{\omega}_{0,k}. \tag{36}$$

This completes the general description of the orthogonalization and cross-fitting procedures. In the coming paragraphs we define the models of interest, and explain how they are related to the procedures described in this section.

3.3.1.1 Partial Linear Regression

In this section we define the main model coming from the DML method, namely the PLR model. This model is defined as

$$y_i=d_i\omega_0+g_0\big(\boldsymbol{x}_i\big)+A_i, \qquad \mathbb{E}\big[A_i|\boldsymbol{x}_i,d_i\big]=0, \tag{37}$$

$$d_i=m_0\big(\boldsymbol{x}_i\big)+B_i, \qquad \mathbb{E}\big[B_i|\boldsymbol{x}_i\big]=0. \tag{38}$$

The corresponding score function filled in with the arguments taken from Equations 37 and 38 is defined as

$$\boldsymbol{\zeta}\big(W_i;\omega_0,\mu_0\big)=\big(y_i-d_i\omega_0-g_0(\boldsymbol{x}_i)\big)\big(d_i-m_0(\boldsymbol{x}_i)\big), \tag{39}$$

where $W_i=\big(y_i,d_i,\boldsymbol{x}_i\big)$ and $\mu=\big(g_0(\cdot),m_0(\cdot)\big)$. Functions $g_0(\cdot)$ and $m_0(\cdot)$ are P-square-integrable, and map the support of $\boldsymbol{x}_i$ to $\mathbb{R}$. Cross-fitting is applied to the PLR model, where the score function used within the cross-fitting procedure is as in Equation (39). Note that this score function is for individual i, these individual score functions are used to calculate the empirical expectations in the cross-fitting procedure.

3.3.1.2 Partial Linear Instrumental Variables

As already explained, endogeneity of the treatment variable is the most common problem in causal inference. Chernozhukov et al. (2018) show that the PLR model can be modified to allow for IV estimation. As the IV method is also researched by us, investigating the PLIV method may give an interesting direct link between the econometric and the ML estimators which aim to solve the same problem. Besides, we can also conclude which type of estimation is more

resistant to outliers, meaning that one of the two types may in general be a more preferred direction of methods to build upon. The PLIV model is defined as

$$y_i = d_i\omega_0 + g_0\big(\boldsymbol{x}_i\big) + A_i, \qquad \mathbb{E}\big[A|\boldsymbol{x}_i, z_i\big] = 0, \tag{40}$$

$$z_i = m_0\big(\boldsymbol{x}_i\big) + B_i, \qquad \mathbb{E}\big[B_i|\boldsymbol{x}_i\big] = 0. \tag{41}$$

The score function corresponding to this model is defined as

$$\boldsymbol{\zeta}\big(W_i; \omega_0, \mu_0\big) = \big(y_i - d_i\omega_0 - g_0(\boldsymbol{x}_i)\big)\big(z_i - m_0(\boldsymbol{x}_i)\big), \tag{42}$$

where $W_i = \big(y_i, d_i, \boldsymbol{x}_i, z_i\big)$ and $\mu = \big(g_0(\cdot), m_0(\cdot)\big)$. Functions $g_0(\cdot)$ and $m_0(\cdot)$ are defined similar to those used in the PLR model.

As for the ML method used for estimation, we make use of post-lasso (Belloni et al., 2013). There is no method which stands out performance-wise, which means that any other suitable method may be chosen. The post-lasso method is more attractive than others when it comes to computation time, which is of importance as we run an extensive simulation study. We use $K = 2$ for the cross-fitting procedure described in Section 3.3.1. In Chernozhukov et al. (2018) it is shown that choosing a larger value for K does not necessarily improve performance, and to limit the computation time a low value for K is the convenient choice.

3.3.2 Approximate Residual Balancing

The ARB method (Athey et al., 2018) also makes use of debiasing, it is therefore not surprising that it is linked to the DML method. The ARB method is however more widely applicable, as it relaxes an assumption made by the DML method. Specifically, the DML method requires consistent estimation of the conditional probability of receiving treatment given the features. Given a linear model, the ARB method relaxes this assumption. In short, it combines weighting with regression, two techniques which are generally used for treatment effect estimation. We refer to Chapter 25.4 and the chapters on regression in Cameron and Trivedi (2005) for a detailed explanation of weighting and regression respectively. For high-dimensional problems, the performances of these techniques fall short when they are separately applied. Athey et al. (2018) show that a combination of both techniques yields better estimators.

Pursuing the approach based on weighting, calculation of the weights typically involves propensity scores. Inaccuracies in propensity score estimates therefore greatly impact the weights. As estimates become poorer as the dimension increases, this approach is inappropriate for high-dimensional problems. The regression-based approach on the other hand, falls short when the propensity scores are not sparse. Combining both techniques as is done in the ARB method overcomes both limitations.

The estimand of the ARB method is defined as

$$\tau = \mu_{\mathrm{T}} - \mu_{\mathrm{C}}, \tag{43}$$

where $\mu_i = \bar{\boldsymbol{x}}_{\mathrm{T}}' \boldsymbol{\beta}_i$ with $\boldsymbol{\beta}_i$ denoting a vector capturing the parameters. This condition holds for $i \in \{\mathrm{T}, \mathrm{C}\}$, where T and C stand for the treatment and control groups respectively. In order to obtain $\boldsymbol{x}_{\mathrm{T}}$, the covariates are averaged over all treated individuals, that is $\bar{\boldsymbol{x}}_{\mathrm{T}} = \frac{1}{n_{\mathrm{T}}} \sum_{\{i:d_i=1\}} \dot{\boldsymbol{x}}_i$, where n_{T} denotes the amount of treated individuals and $\dot{\boldsymbol{x}}_i = \big(x_{i1}, \ldots, x_{ik}, d_i\big)'$. The amount of untreated individuals is logically given by n_{C}. An unbiased estimator of μ_{T} is given as

$$\widehat{\mu}_{\mathrm{T}} = \bar{y}_{\mathrm{T}} = \frac{1}{n_{\mathrm{T}}} \sum_{\{i:d_i=1\}} y_i. \tag{44}$$

Obtaining an estimate of μ_{C} is however more difficult and involves multiple steps.

The procedure of the ARB method for obtaining ATEs is as follows: First, compute positive approximately balancing weights Υ by solving

$$\begin{aligned} \Upsilon = \arg\min_{\widetilde{\Upsilon}} & \Big\{ (1-\iota) \big\| \widetilde{\Upsilon} \big\|_2^2 + \iota \big\| \bar{\boldsymbol{x}}' - \boldsymbol{X}_{\mathrm{T}}' \widetilde{\Upsilon} \big\|_\infty^2 \Big\} \\ \text{s.t.} & \sum_{\{i:d_i=1\}} \widetilde{\Upsilon}_i = 1, \qquad 0 \leq \widetilde{\Upsilon}_i \leq n_{\mathrm{T}}^{-2/3}, \end{aligned} \tag{45}$$

where $\boldsymbol{X}_{\mathrm{T}} = \big(\dot{\boldsymbol{x}}_1, \ldots, \dot{\boldsymbol{x}}_n\big)'$, $\bar{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^n \dot{\boldsymbol{x}}_i$ and $\iota \in \big(0, 1\big)$ is a tuning parameter. Then, fit $\boldsymbol{\beta}_{\mathrm{C}}$ by running an elastic net or lasso regression, that is

$$\widehat{\boldsymbol{\beta}}_{\mathrm{C}} = \arg\min_{\check{\boldsymbol{\beta}}_{\mathrm{C}}} \left[\sum_{\{i:d_i=0\}} \big(y_i - \check{\boldsymbol{x}}_i' \check{\boldsymbol{\beta}}_{\mathrm{C}} \big)^2 + \lambda \Big((1-\alpha) \big\| \check{\boldsymbol{\beta}}_{\mathrm{C}} \big\|_2^2 + \alpha \big\| \check{\boldsymbol{\beta}}_{\mathrm{C}} \big\|_1 \Big) \right], \tag{46}$$

where $\alpha \in \big(0, 1\big]$ and $\lambda > 0$ are tuning parameters.

Finally, we balance the covariates and apply the weights to the residuals, yielding

$$\widehat{\tau} = \bar{y}_{\mathrm{T}} - \left(\bar{\boldsymbol{x}}_{\mathrm{T}}' \widehat{\boldsymbol{\beta}}_{\mathrm{C}} + \sum_{\{i:d_i=0\}} \Upsilon_i \Big(y_i - \dot{\boldsymbol{x}}_i' \widehat{\boldsymbol{\beta}}_{\mathrm{C}} \Big) \right). \tag{47}$$

For our research, we follow the advice of Athey et al. (2018) and set $\alpha = 0.9$ and $\iota = 0.5$, λ is determined by cross-validation.

4 Simulation Study

In this section we describe the simulation study which we conduct for all methods. The goal of our research is to investigate whether the treatment evaluation methods we investigate are robust or not. For our research, it is of interest to evaluate the performances when there are outliers and missing values present in the data. Another way of determining whether a method

is robust or not is by checking a method's performance when one or more of the underlying assumptions are not met. Such a check is also of added value, but as we compare multiple methods this is not convenient. The assumptions differ per method, which leads to too much assumptions which can be relaxed or not. We hence evaluate all methods' performances for the case where the underlying assumptions are met. Investigating performances across methods when the data is contaminated is much easier as the only difference compared to a regular problem lies in the data.

For our simulation study we make use of $S = 100$ runs which enables us to make claims about the methods and the behavior of the corresponding estimators. We drop the subscript of the simulation run in the remainder of this section. Occasionally, we include the subscript of the simulation run in definitions, but only for cases where the definition becomes vague when it is omitted. We set $n = 350$ and $k = 4$ for all scenarios. Although the ML methods are particularly suited for high-dimensional problems, we have decided not to investigate such cases due to computational limitations. It is however an interesting topic for further research to investigate the performances of the ML and econometric methods in settings with high-dimensional data.

The model specifications are given in Section 4.1, an explanation of how the data is contaminated and subsequently imputed is given in Section 4.2. In Section 4.3 we give a general definition of the nonparametric bootstrap, and specify how it relates to our simulation study. Finally, we define all performance measures used for assessing the performances of all methods in Section 4.4.

4.1 Model Specifications

In this section we define all models used in our simulation study. We start off with defining the Data Generating Process (DGP) for the DiD type of methods, these methods rely on the model as in Equation (1). We draw all controls from a multivariate normal distribution with mean zero such that they are i.i.d., that is $\boldsymbol{x}_i \sim N\big(\mathbf{0}_k, \boldsymbol{\Sigma}_k\big)$. Covariance matrix $\boldsymbol{\Sigma}_k$ has ones on the diagonal, the off-diagonal elements are calculated as $\boldsymbol{\Sigma}_{ij} = (-0.9)^{i+j-2}$, where i and j respectively stand for the row and column numbers of the matrix. Variables $d_i \sim Bin\big(1, 0.5\big)$ and $t_i \sim Bin\big(1, 0.5\big)$ are i.d.d. and separately drawn. The error terms are drawn from a standard normal distribution such that they are i.i.d., that is $\epsilon_i \sim N\big(0, 1\big)$. The true parameters are generated as $\beta_{i-1} = -\big(1.1\big)^i + 2 \cdot \text{sgn}\big((-1.1)^i\big)$ for $i = 1, \ldots, k+4$. The dependent variable is calculated according to Equation (1).

For the IV type of methods, we make use of the model as in Equation (3). The instruments z_i and error terms ν_i are separately drawn from a standard normal distribution such that they are i.i.d., that is $z_i \sim N\big(0, 1\big)$ and $\nu_i \sim N\big(0, 1\big)$. The controls are generated differently compared to the DiD type of methods, we draw the first $k-1$ controls from a multivariate normal distribution such that they are i.i.d., that is $(x_{i1}, \ldots, x_{ik-1})' \sim N\big(\mathbf{0}_{k-1}, \boldsymbol{\Sigma}_{k-1}\big)$. Matrix $\boldsymbol{\Sigma}_{k-1}$ again has ones on the diagonal, but the off-diagonal elements are now given by $\boldsymbol{\Sigma}_{ij} = (-0.9)^{2k-(i+j)}$.

The k'th control is drawn such that it is i.i.d. and separately drawn from a standard normal distribution in order to avoid correlation with the other controls, that is $x_{ik} \sim N\big(0,1\big)$.

We generate d_i based on both x_{ik} and z_i, ensuring a correlation with both of the variables. Variable d_i is generated as

$$d_i = \begin{cases} 1 & \text{if } x_{ik} < \bar{x}_k \text{ for } i = 1, \ldots, \dfrac{n}{2} - 1, \text{ or if } z_i < \bar{z} \text{ for } i = \dfrac{n}{2}, \ldots, n, \\ 0 & \text{otherwise,} \end{cases} \tag{48}$$

where $\bar{x}_k = \frac{1}{(\frac{n}{2}-1)} \sum_{i=1}^{\frac{n}{2}-1} x_{ik}$ and $\bar{z} = \frac{1}{(\frac{n}{2})} \sum_{i=\frac{n}{2}}^{n} z_i$. The parameters are generated according to the same formula used for the DiD type of methods, note that the index now only runs until $k+2$. The dependent variable is calculated according to Equation (3). Control x_{ik} is afterwards omitted for estimation, ensuring a correlation between d_i and the error term as the error term now contains x_{ik}.

Finally, we show how the data is simulated for the ML type of methods. Both methods make use of the same data, and the data generation overlaps with the DiD and IV type of methods. In fact, the model used for these methods equals the one given in Equation (3), meaning that it resembles with the model used for the IV type of methods. The controls and error terms are created similarly as with the DiD type of methods, the parameters are generated just as with the IV type of methods. The treatment variable is created just as with the DiD type of methods, and the dependent variable is calculated according to Equation (3).

4.2 Data Contamination and Imputation

In this section we describe how the data is contaminated and imputed afterwards for all methods. We do vary the percentages of outliers and missing values while we keep the amount of individuals and controls fixed. We give a general description of how the data is contaminated, as this procedure is applied to data across multiple methods. Note however that not all variables across all methods are contaminated, we explain per method in detail how contamination takes place.

For a given dataset, we generate one missing value in $\big(100 \cdot \epsilon_{\text{cont}}\big)\%$ of the observations according to the Missing At Random (MAR) mechanism. We refer to Little and Rubin (2002) for a detailed explanation of all missing data mechanisms used in this paper. Likewise, we generate an outlier for one cell within $\big(100 \cdot \epsilon_{\text{cont}}\big)\%$ of the observations. When we choose to generate vertical outliers, the cell corresponding to the dependent variable is always contaminated. When we generate bad leverage points, one of the covariates is randomly selected and subsequently contaminated. Missing values can in practice be present within a binary variable, we have however chosen not to generate missing values for these variables in our simulation study.

In our simulation study and in causal inference in general, these binary variables contain information about the treatment status. All values of these variables are usually known in practice. As estimation of the treatment effect is already a difficult task by itself, uncertainty about the treatment status would only complicate things even more. Hence, we have chosen to mimic real datasets for these variables, meaning that all variables possessing information about the treatment status are not contaminated.

When we generate vertical outliers, the dependent variable is calculated according to

$$\tilde{y}_i = y_i + 100p_i, \tag{49}$$

where $p_i \sim Bin\big(1, \epsilon_{\text{cont}}\big)$. This means that the dependent variable remains the same in about $\big(100 \cdot (1-\epsilon_{\text{cont}})\big)\%$ of the cases, and is incremented with 100 in the remaining cases. We are only interested in investigating the biases arising from contamination, the way we contaminate the variable is therefore not of great importance. Note that there are numerous ways to contaminate variables, we increment the dependent variable with 100 as it is sufficient for the comparison of different methods in our paper.

The bad leverage points are generated in a different way, as we randomly select a variable to be modified. Let us denote the randomly selected variable as x_{il}, where $l \in \big\{1, \ldots, p\big\}$. If we investigate the DiD and ML type of methods, $p = k$. When we examine the IV type of methods, we set $p = k + 1$ as outliers are also generated in the instrument, we include z_i as x_{ip} in this case for convenience. The variable randomly selected is updated according to

$$\tilde{x}_{il} = x_{il} + 15p_i. \tag{50}$$

The same reasoning of the value chosen to add to the dependent variable to come up with vertical outliers also holds for the generation of bad leverage points.

With the DiD type of methods, we contaminate all non-binary variables. These methods rely on the parallel trends assumption in practice, but vertical outliers can be present after treatment. Meeting the parallel trend assumption excludes the presence of vertical outliers in observations measured over the corresponding time period. The parallel trend assumption gives however no guarantee of the behaviour of the observations after treatment, meaning that the presence of vertical outliers is not excluded after treatment. Bad leverage points can on the other hand be present even when the parallel trend assumption is justified. Hence, generation of vertical outliers and bad leverage points is desirable as they can go unnoticed in practice. Missing values are also created for all non-binary variables, all outliers and missing values are generated according to the procedure described earlier in this section.

With the IV type of methods, variables are contaminated similarly as with the DiD type of methods. Additionally, missing values and outliers are also generated for the instrument. The vertical outliers are generated as described in the general procedure, the generation of bad leverage points differs slightly for the IV type of methods. The instruments we generate in our simulation study are continuous, as they are continuous in most of the real datasets. In such cases, outliers are as likely to be present in both the instruments and controls. Hence, when we generate bad leverage points, either one of the cells of the controls or the cell of the instrument is contaminated with an outlier. As the endogenous variable is fitted by making use of the instruments within the first stage of IV estimation, outliers in the instrument cause bad leverage points. Simultaneously generating vertical outliers and outliers in the instrument is undesirable, as this can lead to good leverage points which in turn do not bias the parameter estimates.

Finally, all non-binary variables are contaminated with both outliers and missing values within the ML type of methods.

If the missing values in a dataset are Missing Not At Random (MNAR) or MAR, omission of missing values corrupts the data. Missing values in real data often follow the MAR mechanism, which is why we have chosen to generate the missing values according to this mechanism as well. Instead of deletion, imputation is the better way of dealing with missing values. Numerous imputation techniques exist, the appropriateness of a technique however depends on the data. As we generate correlated data, in turn drawn from a distribution, model-based imputation can be expected to perform well. Model-based imputation is however computationally expensive, which makes it unsuitable for our research. Instead, we make use of k Nearest Neighbor (kNN) imputation (Troyanskaya et al., 2001), this technique is computationally less expensive, and using the median for aggregation makes it robust to outliers.

As we make use of imputation to handle the missing values, standard inference which is based on a fully observed data matrix is not valid anymore. We have to take extra uncertainty into account as some elements of the imputed data matrix are estimated. One way of incorporating such additional uncertainty is by applying the bootstrap (Efron, 1992,9). The bootstrap approximates an estimator's distribution by sampling with replacement from the observed data. The bootstrap works if the asymptotic distribution of an estimator is normal (Mammen, 2012). We can also utilize the bootstrap to simulate the missing data mechanism prior to imputing the missing values (Efron, 1994), resulting in the nonparametric bootstrap. We define the nonparametric bootstrap in the next section, and subsequently explain when the method works properly.

Finally, we would like to make a note on which observations are used for estimation. Although we contaminate the dependent variable with outliers and missing values, it is not desirable to use all of these observations for estimation. We make use of the approach opted in Von Hippel (2007), in this paper the idea is proposed to delete all observations with imputed dependent variables after the imputation step. Due to this strategy, additional information regarding the

known values of these observations is used for the imputation of missing values in the covariates. Omitting these observations prior to imputation would result in a loss of information. On the other hand, keeping the observations with imputed values for the dependent variable is not desirable. As these values are unknown, it comes down to regressing a prediction on the covariates. As uncertainty arises from imputation, such an approach is likely to bias the parameter estimates.

We investigate two different contamination scenarios in our simulation study, we set $\epsilon_{\text{cont}} = 7.5\%$ for the first scenario. Due to the nature of our cellwise contamination, an observation containing a cellwise outlier automatically becomes a rowwise outlier. Due to the programmed nature of the simulation, the total contamination ranges from 7.5% to 15%, as missing values and cellwise outliers may or may not simultaneously occur within an observation. For the second scenario we set $\epsilon_{\text{cont}} = 20\%$, meaning that the amount of contaminated observations lies between 20% to 40%.

4.3 Nonparametric Bootstrap

For R bootstrap replications, the nonparametric bootstrap is defined as follows:

For $r = 1, \ldots, R$, generate bootstrap sample $\left(\boldsymbol{x}_{1r}^*, \ldots, \boldsymbol{x}_{nr}^*\right)'$ by sampling with replacement from the data containing missing values, that is $\left(\boldsymbol{x}_{1\text{miss}}, \ldots, \boldsymbol{x}_{n\text{miss}}\right)'$ where $\boldsymbol{x}_{j\text{miss}}$ contains all variables of a single observation for $j = 1, \ldots, n$. Next, impute missing values in $\boldsymbol{X}_r^* = \left(\boldsymbol{x}_{1r}^*, \ldots, \boldsymbol{x}_{nr}^*\right)'$ to obtain imputed matrix $\hat{\boldsymbol{X}}_r^*$. Finally, compute bootstrap replicate $T_r^* = T\left(\hat{\boldsymbol{X}}_r^*\right)$ and store this replicate for further calculations.

After calculating all replications, parameter estimates are obtained by averaging over all replications, that is $\bar{T}^* = \frac{1}{R}\sum_{r=1}^{R} T_r^*$. Standard errors are also calculated based on all bootstrap replications, that is $\widehat{\sigma}_{\bar{T}^*} = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}\left(T_r^* - \bar{T}^*\right)^2}$. There is debate about what number of bootstrap replications is sufficient, the amount of replications needed namely depends on the data. In most of the problems, one thousand replications show to approximate the estimator's distribution. A larger number of replications is always better, due to computational limitations we however have to make a trade off. In order to limit the computation time, we follow this number accepted by the majority of the researchers and set $R = 1000$.

As already mentioned, we make use of the nonparametric bootstrap, but the parametric bootstrap logically also exists. The difference between the appropriateness of both models relies on the missing data mechanism. For data which is MAR or Missing Completely At Random (MCAR), the nonparametric bootstrap works properly. If the data is MNAR, the parametric bootstrap has to be used. If the data is MAR or MCAR, nonresponse is ignorable, meaning that no additional problems arise during estimation. If the data is MNAR, nonrepsonse is not ignorable, meaning that additional data analysis has to be performed before the method which was intentionally meant to be used can be employed. In particular, this means that the

missing data mechanism should be explicitly modelled. For our simulation we however generate MAR data, as this resembles with the majority of the missing data mechanisms in real datasets.

There are only a few conditions which must be satisfied for the nonparametric bootstrap to work properly, which makes it a popular method in practice. The main disadvantage of this easily applicable method is its computation time, which can grow large compared to other methods. Another imputation method which also corrects for uncertainty regarding the estimations of the missing values is the multiple imputation method (Rubin, 2004). This is the go to method if computation times are wished to be small, but it imposes stronger assumptions. For example, an estimator's variance has to be calculable within a bootstrap replication.

As we investigate the RIV method as explained in Section 3.2.2, using the multiple imputation method is inappropriate. In Freue et al. (2013), formulas of the standard errors are derived for the S-estimator, these standard errors are however asymptotic. As convergence of the standard errors depends on the number of observations, we have chosen to employ the bootstrap. Standard errors obtained from this method are more accurate when the sample size is low, and therefore more reliable in general. Although most of the modern datasets contain a large amount of observations, datasets with a small amount of observations can still be encountered.

In order to link the nonparametric bootstrap to our simulation study, we let $\widehat{\beta}_{isr}$ denote the estimator of the i'th parameter in replication r of simulation run s. At the end of every simulation run, a parameter estimate is calculated by averaging over all replications, that is

$$\widehat{\beta}_{is} = \frac{1}{R}\sum_{r=1}^{R}\widehat{\beta}_{isr}. \tag{51}$$

Final parameter estimates are in turn determined by averaging over all simulation runs, that is

$$\widehat{\beta}_{i} = \frac{1}{S}\sum_{s=1}^{S}\widehat{\beta}_{is}. \tag{52}$$

Standard errors are estimated based on the parameter estimates retrieved from all simulation runs, that is

$$\widehat{\sigma}_{\beta_i} = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\widehat{\beta}_{is} - \widehat{\beta}_{i}\right)^2}. \tag{53}$$

In order to make claims about a method's performance we have to make use of more metrics than just the estimated parameters and standard errors. The performance measures on which we base the methods' performances in our simulation study are defined in Section 4.4.

4.4 Performance Measures

In this section we define all metrics used to evaluate the performances of the methods investigated. Besides metrics that assess the quality of estimates obtained from regression techniques, we also define measures used for determining imputation quality. Starting off, we make use of the Root Mean Squared Error (RMSE), an error measurement which assigns higher weights to estimates further away from the true parameter. As we are investigating causality, accurate point estimates are of big importance, making the RMSE a suitable metric. The RMSE averaged over all parameter estimates is calculated as

$$RMSE = \sqrt{\frac{1}{pS}\sum_{i=1}^{p}\sum_{s=1}^{S}\left(\beta_i - \widehat{\beta}_{is}\right)^2}, \tag{54}$$

where p denotes the amount of parameters estimated in a model. An intuitive explanation of the RMSE as given in Equation (54) is that we average the Mean Squared Errors (MSEs) of all parameter estimates in every simulation run. Eventually, we also average these MSEs over all parameters to end up with a single number, which simplifies the comparison of RMSE scores across different methods.

As the MSE can be written as a metric consisting of the bias and variance, a high MSE value can be due to multiple reasons. Specifically, the MSE equals the sum of the squared bias and the variance. A high MSE can therefore be due to a high squared bias, a high variance or both. Measuring at least one of these components sheds light on the MSE, in this paper we only measure the bias. The bias averaged over all parameter estimates is calculated as

$$Bias = \frac{1}{pS}\sum_{i=1}^{p}\sum_{s=1}^{S}\left(\beta_i - \widehat{\beta}_{is}\right). \tag{55}$$

In order to determine standard error accuracy, we make use of the coverage defined as

$$Coverage = \frac{1}{pS}\sum_{i=1}^{p}\sum_{s=1}^{S} I\left[\hat{\beta}_{is} - t_*\hat{\sigma}_{\beta_i} \leq \beta_i \leq \hat{\beta}_{is} + t_*\hat{\sigma}_{\beta_i}\right], \tag{56}$$

where t_* is the distribution's critical value for a certain significance level. For a two-tailed distribution, $t_* = t_{\alpha/2}$, where we set significance level $\alpha = 0.05$. We follow the same strategy as for the RMSE, meaning that we eventually average coverages over all parameters and simulation runs to end up with a single number. Standard errors are found to be accurate if the coverage approximately equals $(1 - \alpha)$, indicating that the amount of times which the parameters fall within the confidence interval is as expected. This conclusion only holds if the parameters do not vary to a great extent across simulation runs, which is in turn true if the variance component of the RMSE is found to be low.

We also investigate the predictive performances of all methods. For example within healthcare, it may be of interest to predict what the effect of treating a patient will be on an outcome variable of interest. We omit the subscript of the bootstrap replication for convenience. A logical consequence of the bootstrap is that a fraction of the observations do not appear in the bootstrap sample. As both a training and test set are needed for a predictive analysis, we can easily construct the test set by assembling the observations which do not appear in the bootstrap sample. As explained in Efron and Tibshirani (1997), the training error is defined as

$$\overline{err} = \frac{1}{n}\sum_{i=1}^{n} L\big(y_i, f(\ddot{\boldsymbol{x}}_i)\big), \tag{57}$$

where $L(\cdot)$ is a loss function, $\ddot{\boldsymbol{x}}_i$ contains all covariates.

We use the MSE as the loss function, meaning that $L\big(y_i, f(\ddot{\boldsymbol{x}}_i)\big) = \big(y_i - f(\ddot{\boldsymbol{x}}_i)\big)^2$. Solely using the training error is discouraged, as it is downward biased. Predicting such a training set by making use of a model fit to multiple bootstrap samples is a naive strategy, as they have observations in common. We apply the leave-one-out bootstrap, where the out-of-sample error is defined as

$$\widehat{Err}_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\big|I^{-i}\big|} \sum_{b \in I^{-i}} L\big(y_i, f_b(\ddot{\boldsymbol{x}}_i)\big), \tag{58}$$

where I^{-i} is the set of indices of bootstrap samples that do not contain observation i. If all samples contain a certain observation, we leave steps concerning that observation out of the calculation. This measurement is however upward biased, as some observations occur more than once in a bootstrap sample. As a reoccuring observation does not provide as much information as a new one, the error term is overestimated.

To alleviate these biases, Efron (1983) proposes the .632 bootstrap which weighs the training and out-of-sample errors, therefore trying to find a balance between the down and upward biases. When the predictions however overfit the data, that is $\overline{err} = 0$, the .632 bootstrap will underestimate the prediction error. The .632+ bootstrap proposed by Efron and Tibshirani (1997) is based on the .632 bootstrap, but additionally measures the degree of overfitting. The corresponding error measurement is defined as

$$\widehat{Err}_{.632+} = \big(1 - w\big)\overline{err} + w\widehat{Err}_1, \tag{59}$$

where

$$w = \frac{0.632}{1 - 0.368R}. \tag{60}$$

The degree of overfitting is in turn given by

$$R = \frac{\widehat{Err}_1 - \overline{err}}{\gamma - \overline{err}}, \tag{61}$$

where γ is the no-information error rate, that is

$$\gamma = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} L\big(y_i, f(\ddot{\boldsymbol{x}}_j)\big). \tag{62}$$

The intuition behind Equation (61) is that the degree of overfitting increases if the difference between the training and out-of-sample errors increases, meaning that the predictions of the training error correspond too closely to the data. As the degree of overfitting increases, a higher weight is assigned to the out-of-sample error. If there is no overfitting, $\widehat{Err}_1 = \overline{err}$, which in turn yields $R = 0$. This reduces to the .632 bootstrap, that is Equation (59) with $w = 0.632$.

Eventually, we average the errors over all simulation runs to come up with the final prediction error. For convenience, we do not define a mathematical formula for this error measure as it is likely to come over as confusing. In short, the final prediction error is calculated by averaging over the .632+ bootstrap errors obtained from all simulations, where the .632+ bootstrap error is as in Equation (59). We use this final error measurement to compare all methods' predictive performances, the method corresponding to the minimal value excels at predicting outcomes for new data.

As the outliers are simulated, we can pinpoint the exact location of these values within the dataset. As both the IV-DDC and RIV methods flag outliers, we can evaluate both methods' outlier detection accuracies. These methods however differ within this detection, as the IV-DDC method flags cells as outliers. The RIV method flags observations as outliers, meaning that we need separate measurements for both methods. In our simulation, we however generate one outlier per observation, if an observation is selected for contamination. We determine outlier detection accuracy of the RIV method by checking whether the observation flagged as outlying contains an outlying cell or not. This way, the outlier detection abilities of both the IV-DDC and RIV methods can be compared one on one.

Starting off with the IV-DDC, we evaluate the outlier detection performance as

$$OutAccCellwise = \frac{1}{n_{\mathrm{o}}} \sum_{i=1}^{n} \sum_{j=1}^{p+1} Out_{ij}, \tag{63}$$

where

$$Out_{ij} = \begin{cases} 1 & \text{if } \tilde{x}_{ij} \text{ is an outlier and correctly detected,} \\ 0 & \text{otherwise.} \end{cases} \tag{64}$$

With n_{o} we denote the number of outlying cells present in the data and $\tilde{x}_{ij}$ is as in Section 4.2, note that we include $\tilde{y}_i$ as $\tilde{x}_{ip+1}$ for convenience. We generalize this metric to a measurement for evaluating how many observations are correctly detected as outlying by the RIV method,

that is

$$OutAcc = \frac{1}{n_{\text{o}}} \sum_{i=1}^{n} Out_i, \tag{65}$$

where

$$Out_i = \begin{cases} 1 & \text{if } (\tilde{x}_{i1}, \ldots, \tilde{x}_{ip+1})' \text{ is detected as outlying and contains an outlying cell,} \\ 0 & \text{otherwise,} \end{cases} \tag{66}$$

In the end, we average both the metrics in Equations 63 and 65 over all simulation runs, just as with the final prediction error, to end up with the final imputation errors.

Finally, we define metrics used for assessing the quality of the missing value imputations. For these steps we make use of the metrics proposed in Templ et al. (2011), which distinguish between error measurements for categorical, binary, continuous and semi-continuous variables. We deal with binary and continuous variables in our simulation, but we only generate missing values for the continuous variables. Hence, we adjust the metric proposed in Templ et al. (2011) and measure the Absolute Relative Error (ARE) of the imputations for the continuous variables as

$$ARE_{\text{imp}} = \frac{1}{n_{\text{miss}}} \sum_{i=1}^{n} \sum_{j=1}^{p+1} \left| \frac{\tilde{x}_{ij}^{\text{orig}} - \tilde{x}_{ij}^{\text{imp}}}{\tilde{x}_{ij}^{\text{orig}}} \right|, \tag{67}$$

where $\tilde{x}_{ij}^{\text{orig}}$ and $\tilde{x}_{ij}^{\text{imp}}$ stand for the original and imputed value respectively, n_{miss} stands for the amount of missing values. A minimal value of the error measurement indicates an imputation lying closer to the true value on average.

A disadvantage of the ARE metric is that it is likely to take on high values as the original values approach zero. As we draw data from a multivariate normal distribution with zero mean and relatively low covariances, values close to zero are likely to be present. Hence, we also measure the Mean Absolute Error (MAE) of the imputations, that is

$$MAE_{\text{imp}} = \frac{1}{n_{\text{miss}}} \sum_{i=1}^{n} \sum_{j=1}^{p+1} \left| \tilde{x}_{ij}^{\text{orig}} - \tilde{x}_{ij}^{\text{imp}} \right|. \tag{68}$$

The MAE measures how much an imputation differs from the original value on average, without taking the magnitudes of both values into account. Logically, just as with the ARE, a lower value of the MAE is desirable. In cases where the ARE yields a large error due to an original value lying close to zero, the MAE does a better job at capturing the quality of the imputation. Finally, just as with the prediction and outlier detection measures, we average the AREs and MAEs obtained from all simulation runs to end up with the final metrics of imputation quality.

5 Real Data

In this section we describe the two real datasets to which some of the methods investigated in this paper are applied to. The results which we replicate and extend are not based on explanatory variables including missing values, meaning that making use of the imputation strategy as given in Section 4.2 is superfluous. We do however check all datasets for outliers, and pay specific interest to the outcomes of the proposed robust alternatives. An explanation of the real dataset used for an application of the IV type of methods described in this paper is given in Section 5.1. The second real dataset is used for an application of the DiD type of methods, it is described in Section 5.2.

5.1 Dinkelman (2011)

For the IV type of methods we make use of the data regarding electrification and its effect on employment in South Africa (Dinkelman, 2011). Results show that electrification causes a significant increase in the female employment rate within a five year period. Besides this main finding, other interesting relationships regarding male and female employment are obtained and discussed. The majority of the South African population did not have access to electricity prior to the elections of 1994. Based on the election outcomes, the South African government decided to increase provisioning of such basic services. From 1995 onwards, a set amount of South African households is annually provided with electricity.

An IV approach might seem unnecessary at first glance, as random assignment of electricity seems feasible. Motivations behind this non-random assignment are given in Dinkelman (2011), the non-random selection of communities which were to receive electricity was mainly due to higher powers of politicians. Hence, the treatment variable indicating whether a household has received electricity on behalf of the project set up by the government is endogenous. This variable is instrumented by a measure of land gradient, the choice of this instrument is based on the fact that a higher gradient increases the costs of the household's electrification, which means that it plays a role in electricity assignment. In Dinkelman (2011) it is shown that this instrument is indeed valid, leading to a proper application of the IV method.

All variables used in Dinkelman (2011) are continuous except for the binary treatment variable. The number of controls used varies between 10 and 12, meaning that outliers in one or more of these controls are possibly present. As we illustrate in Section 6.2, some variables are large in magnitude, meaning that an extreme data point is likely to bias the results if the effect of this variable on the dependent variable is significant and relatively large. A detailed analysis of our replication and extension of the results of Dinkelman (2011) is given in Section 6.2.

5.2 Card and Krueger (1993)

For the DiD type of methods we make use of the data originally used in Card and Krueger (1993) to study the effects of an increase in minimum wage. This paper is a famous example of the application of the DiD method, as it contradicts a pattern which was believed to hold based on theoretical economic work. The main finding of Card and Krueger (1993) is that an increase in the minimum wage leads to an increased level of employment. As mentioned earlier, such a wage inrease was suspected to decrease the level of employment before this paper was published. As this conclusion was striking to a lot of researchers, further research on this topic gained a lot of interest.

In general, a potential problem for research conducted is using data which does not accurately represent the problem. The sample used in Card and Krueger (1993) is relatively small, which calls the sample's representativeness of the population into question. In Neumark and Wascher (2000) it is claimed that the results obtained from Card and Krueger (1993) paint a wrong picture, as they obtain a negative relation between a minimum wage increase and the level of employment by using different data. What makes it particularly interesting, is that Neumark and Wascher (2000) make use of data regarding restaurants from the same food chains. As these restaurants are also located in New Jersey and Pennsylvania, the potential problem of differing characteristics is unlikely to be present in this case.

In a reply to Neumark and Wascher (2000), Card and Krueger (2000) further investigate the issue. They refute the conclusion drawn in Neumark and Wascher (2000), but dig deeper into the problem as they now also doubt the credibility of the conclusion drawn in Card and Krueger (1993). After investigating multiple datasets of which the appropriateness is illustrated, Card and Krueger (2000) conclude that the increase in minimum wage does not have a significant effect on the level of employment. Even though the original authors have already rejected their claim themselves, a search for potential outliers has not yet been carried out. In Card and Krueger (2000) they do highlight a couple of stores which show behaviour differing from the majority, but using the RDiD method provides us with more information. At the time, robust regression was not as widely used as it is today, we are therefore interested in applying our proposed robust alternative of the DiD method to the data.

6 Results

In this section the results obtained from both the simulation study, as well as from the real data are given. In Section 6.1 we discuss the results obtained from the simulation study described in Section 4, where we investigate multiple cases which differ in the extent to which they are contaminated. In Section 6.2 we investigate the results obtained from analyzing the two real datasets described in Section 5.

6.1 Simulation Study

We start off with investigating the DiD type of methods. The results obtained from the DiD type of methods for $\epsilon_{\text{cont}} = 7.5\%$, where outliers are caused due to bad leverage points are given in Table 1.

Table 1: The results of the DiD type of methods for $\epsilon_{\text{cont}} = 7.5\%$ and outliers due to bad leverage points.

	RMSE	Bias	Coverage	ATE		Prediction	Outlier Detection	Imputation ARE	Imputation MAE
DiD	1.609	0.097	69.125%	4.223	(1.824)	39.016	-	2.541	0.481
RDiD	**0.155**	**0.024**	**93.625%**	**4.143**	(0.231)	**1.027**	0.999		

Notes: The standard errors are given in parentheses.

Table 1 shows the most important results from both the DiD type of methods, where the numbers in bold correspond to the best performing method for a certain metric. We see superior performances for the RDiD method based on all metrics given in Table 1, but we will subsequently walk through all findings. The RMSE is denoted in the first column, which is minimal for the RDiD method. As both biases are low, this indicates that the parameter estimates of the RDiD method differ less on average from the true parameters than for the DiD method. The coverage of the RDiD method is close to the in this case ideal value of 95%, the coverage of the DiD method lies further away. The coverage of the RDiD method does not equal 95%, but there is a logical explanation for this.

As we make use of the bootstrap, coverage deviations are either due to a too small number of bootstrap replications, a too small number of simulation runs or both. The deviations are however small, and due to computational limitations we do not investigate these cases for a larger amount of bootstrap replicates and/or simulation runs. If we were to judge the standard errors on these outcomes, we would argue only the standard errors of the RDiD method to be accurate. The coverage however also depends on the parameter estimates, as the method relies on calculating confidence intervals. The coverage may lead to wrong conclusions if the parameter estimates vary considerably, as the confidence intervals become unrealistically large. Combining the RMSE and the bias of the RDiD method, which gives us information about the variance, shows that the variance is low. Hence, the corresponding parameter estimates do not vary to a great extent, the standard errors are therefore accurate.

The ATEs of both methods lie close to the true value of 4.145, the corresponding standard error of the RDiD method is however much smaller than that obtained from the DiD method. As there is more uncertainty regarding the ATE obtained from the DiD method, the RDiD method is superior in this case. The prediction error corresponding to the RDiD method is a lot smaller than that of the DiD method, meaning that it excels at predicting outcomes for new data. The superiority of the RDiD method can easily be explained, as 99.9% of the outliers are on average correctly detected by the method. In this simulation setting, the outlying distribution differs from the population distribution to a great extent, making it relatively easy for the

RDiD method to detect the outliers. Once this difference shrinks, we expect the percentage of correctly detected outliers to decrease as well. Due to computational limitations, we leave this for further research.

The ARE of the imputations shows that imputations on average deviate from the original value by 254.1%, meaning that the missing values are imputed rather poorly. As noted in Section 4.4, the ARE may be misleading when values lie close to zero. The MAE shows that imputations on average deviate from the original value by a value of 0.481, meaning that the imputations are not as poor as indicated by the ARE. Still, we demonstrate that kNN imputation is only able to mimic the patterns in the original dataset to a certain extent.

The results of the IV type of methods under the same setting as the DiD type of methods are given in Table 2. As already mentioned, the instrument can and therefore is also contaminated with outliers and missing values in our case. A method which breaks down when outliers are present in the instrument may show promising results within a simulation study for a dataset with a clean instrument.

Table 2: The results of the IV type of methods for $\epsilon_{\text{cont}} = 7.5\%$ and outliers due to bad leverage points.

	RMSE	Bias	Coverage	ATE		Prediction	Outlier Detection	Imputation ARE	Imputation MAE
IV	3.447	-0.590	63.400%	2.126	(6.121)	178.640	-	2.697	0.613
RIV	**0.592**	**0.026**	94.400%	**3.962**	(1.055)	**157.860**	0.793		
IV-DDC	4.864	-0.532	64.200%	2.273	(9.316)	237.812	**0.799**		
PLIV	8.286	-1.896	**95.000%**	1.876	(8.107)	-	-		

Notes: The standard errors are given in parentheses.

Table 2 shows superiority of the RIV method based on almost all of the measures given in the table. An RMSE of 0.592 for the RIV method indicates that the parameter estimates do not vary a lot between simulation runs. The RMSE and bias values of all other methods however indicate the opposite, meaning that averaging over all simulation runs does not accurately represent all simulation outcomes. Hence, the coverage is a metric which is nonsensical to interpret in this setting for all methods except for the RIV method. For the RIV method, the coverage lies close to the expected 95% mark, indicating accurate standard errors. Although the coverage corresponding to the PLIV method equals 95%, its variance is too high for the coverage to be reliable.

The RIV method does best at estimating the treatment effect of 3.772, with a corresponding standard error which is acceptable. The treatment effect estimates of all other methods also come close to the true treatment effect, due to their large standard errors there is however too much uncertainty about the parameter estimates. Prediction-wise, the RIV method outperforms the other methods, although the differences in performance are rather small in this case. Note that predictions cannot be made for the PLIV method, as this method does not return parameter estimates other than the ATE.

Both the RIV and IV-DDC methods correctly detect almost 80% of the outliers on average. As all outliers are similarly generated, differences between these methods are mainly due to the quality of the imputations, which is something on which the performance of the IV-DDC method highly depends. On average, an imputation differs almost 270% relative to the original value. The MAE is however the metric of greater interest in this case. On average, an imputation has a deviation of 0.613 from the original value, which can be argued to be acceptable. The results in Table 2 however show that setting outliers to missing and imputing the values afterwards performs worse than the IV method. This in turn indicates that the imputations are of a quality which only worsens the results.

The results of the ML type of methods under the same setting as the previously examined methods are given in Table 3. These methods are especially designed for calculating the ATE, all results given in Table 3 are therefore not averaged over multiple parameters. All measurements correspond to the estimated treatment effects obtained over all simulation runs.

Table 3: The results of the ML type of methods for $\epsilon_{\text{cont}} = 7.5\%$ and outliers due to bad leverage points.

	RMSE	Bias	Coverage	ATE		Imputation ARE	Imputation MAE
ARB	1.291	0.100	94.000%	3.871	(1.294)	1.843	0.419
PLR	**1.079**	**0.096**	96.000%	**3.867**	(1.081)		

Notes: The standard errors are given in parentheses.

Table 3 shows similar results for the ARB and PLR methods. Both the RMSEs are relatively small, this indicates that the estimate of the treatment effect accurately captures the estimates from all simulation runs as the biases are low. We do however note that the PLR method obtains a lower RMSE compared to the ARB method, meaning that the estimates obtained from the PLR method vary less across simulations. Subsequently, we can draw conclusions from the coverage values as the parameter estimates are found to be accurate. Both the coverages approach the desirable 95% level, meaning that the standard errors are accurate for both methods.

Both the ATEs are close to the actual treatment effect of 3.772, the standard error associated with the PLR method is however smaller, indicating less uncertainty about the ATE. The ARE corresponding to the imputations again suggests poor imputations, as an imputation deviates 184.3% from its original value on average. The MAE shows the difference between an imputed and the original value to equal 0.419 on average, meaning that the imputations are not as bad as indicated by the ARE. Although the quality of the imputations is controversial, both the PLR and ARB methods show to be robust in this setting.

Next, we investigate all methods in similar settings as the ones previously described. The outliers now however concern vertical outliers instead of bad leverage points. The results of the DiD type of methods are given in Table 4.

Table 4: The results of the DiD type of methods for $\epsilon_{\text{cont}} = 7.5\%$ and outliers due to vertical outliers.

	RMSE	Bias	Coverage	ATE		Prediction	Outlier Detection	Imputation ARE	Imputation MAE
DiD	4.658	0.842	87.375%	3.558	(6.234)	83.858	-	2.547	0.529
RDiD	**0.155**	**0.024**	**94.375%**	**4.121**	(0.224)	**1.028**	0.999		

Notes: The standard errors are given in parentheses.

The results obtained from Table 4 lead to the same conclusions drawn based on Table 1, where the only difference between the datasets is the type of outlier generated. When vertical outliers are present, the DiD method performs worse relative to the RDiD method, as can be seen from the higher RMSE and prediction error values. The results obtained from the DiD method in Table 1 were found to be credible to a certain extent, the ATE and its standard error were somewhat accurate. When vertical outliers are present, the standard error increases so much that the ATE is not even significantly different from zero anymore.

The results of the IV type for $\epsilon_{\text{cont}} = 7.5\%$ when outliers are caused by vertical outliers are given in Table 5.

Table 5: The results of the IV type of methods for $\epsilon_{\text{cont}} = 7.5\%$ and outliers due to vertical outliers.

	RMSE	Bias	Coverage	ATE		Prediction	Outlier Detection	Imputation ARE	Imputation MAE
IV	5.517	1.476	89.400%	3.992	(8.156)	324.355	-	2.687	0.647
RIV	**0.592**	0.016	**95.200%**	**3.867**	(1.057)	159.492	**1.000**		
IV-DDC	0.616	**-0.006**	94.200%	3.561	(1.065)	**159.285**	0.987		
PLIV	8.065	0.256	96.000%	4.027	(8.102)	-	-		

Notes: The standard errors are given in parentheses.

Table 5 shows results which lead to the same conclusions drawn from the setting where outliers were caused due to bad leverage points. We shortly go through the results as the differences between the methods' performances have changed. The RMSE outcomes lead to the same conclusions, we do however note that the RMSE of the IV-DDC method nears that of the RIV method. Bad leverage points detected as outliers by the IV-DDC method were imputed and used within the estimation. A similar strategy is followed with the vertical outliers, these observations are however omitted before estimation, as already explained in Section 4.2. In general, deletion of outliers in a dataset is not a straightforward procedure. Within this simulation it yields favorable results, but it is not a strategy which can always be followed in practice.

Both the coverages of the RIV and IV-DDC methods approach 95%, indicating accurate standard errors. The coverages of the IV and PLIV methods are non-credible as their variances are too large. All ATEs lie close to the true value of 3.772, with the RIV and IV-DDC methods showing plausible standard errors. Prediction-wise, the RIV and IV-DDC methods show performances which are roughly equal in quality. Compared to the RIV and IV-DDC methods, the predictions from the IV method leave a lot to be desired. Bad leverage points were already correctly detected to a large extent, but almost all vertical outliers are detected in this setting.

The RIV method succesfully detects all outliers, while the IV-DDC method correctly detects 98.7% of the outliers.

Next we investigate the ML type of methods for $\epsilon_{\text{cont}} = 7.5\%$ when outliers are caused by vertical outliers, the results obtained are given in Table 6.

Table 6: The results of the ML type of methods for $\epsilon_{\text{cont}} = 7.5\%$ and outliers due to vertical outliers.

	RMSE	Bias	Coverage	ATE		Imputation ARE	Imputation MAE
ARB	3.100	**-0.342**	**95.000%**	**3.429**	(3.097)	2.069	0.482
PLR	**3.081**	-0.367	96.000%	3.404	(3.075)		

Notes: The standard errors are given in parentheses.

Table 6 shows similar, and rather poor results for both methods. Both methods attain high values for the RMSE and standard errors. Both coverages are close to the optimal value, these values however have no meaning as the variances are large. The parameter estimates roughly approach the true value of 3.772 of the treatment effect, but there is too much uncertainty about these estimates due to the large standard errors.

Overall, we see superiority of the RDiD method with respect to the DiD method for $\epsilon_{\text{cont}} = 7.5\%$, making it the favoured method when outliers are present. The RIV method is the best performing one out of all IV type of methods, meaning that the inclusion of a binary variable does not necessarily harm its performance to a great extent. We do however draw all controls from a multivariate normal distribution, meaning that all controls fit the elliptical structure very well. If the controls do not suit an elliptical structure in practice, we do not promote the use of the IV-DDC method if the imputation method performs poorly. Finally, the ARB and DML methods show similar performances. They are robust to bad leverage points, but the performances visibly worsen in the presence of vertical outliers.

We also examine all methods for $\epsilon_{\text{cont}} = 20\%$, but the results lead to conclusions similar to those drawn from the results of $\epsilon_{\text{cont}} = 7.5\%$. As more data is contaminated, the results are more extreme, but the order of methods' performances and their superiority does not change. These results can be found in Tables 12 to 17 in Section A.

6.2 Real data

In this section we examine the two real datasets described earlier. The analysis of the dataset used in Card and Krueger (1993) is given in Section 6.2.1, the results obtained from the data used in Dinkelman (2011) are given in Section 6.2.2.

6.2.1 Card and Krueger (1993)

For investigating the real data as described in Section 5, we start off with applying the DiD and RDiD methods to the data used in Card and Krueger (1993). Except for in the dependent variable, there are no values missing. Hence, we do not make use of the imputation strategy as described in Section 4.2 as we would eventually delete all observations with imputations. Rather, we omit all stores with at least one missing independent variable, meaning that we are left with $n = 384$ stores where the employment level is known both pre and post treatment. The model to be estimated is defined as

$$emp_{it} = \beta_{0,\text{emp}} + \beta_{1,\text{emp}} D_t + \beta_{2,\text{emp}} NJ_i + \beta_{3,\text{emp}}\big(D_t \cdot NJ_i\big) + \epsilon_{it}, \tag{69}$$

where $t = 0$ and $t = 1$ indicate observations pre and post treatment respectively. Variable emp_{it} stands for the average employment of store i at time t, D_{it} equals one if $t = 1$ and zero otherwise. Finally, NJ_i equals one if a store is located in New jersey and zero if a store is located in Pennsylvania. The parameter of interest is $\beta_{3,\text{emp}}$, which represents the treatment effect.

As stated in Cameron et al. (2008), clustered robust standard errors fall short in this case as there are only two states. Hence, we apply the wild cluster bootstrap as also described in Cameron et al. (2008). The results obtained are given in Table 7 for $R = 1000$.

Table 7: The results of the DiD and RDiD methods applied to the data used in Card and Krueger (1993).

	DiD	RDiD
$\widehat{\beta}_{0,\text{emp}}$	23.42***	21.49***
	(1.19)	(1.13)
$\widehat{\beta}_{1,\text{emp}}$	-2.32	-0.95
	(1.66)	(1.49)
$\widehat{\beta}_{2,\text{emp}}$	-3.01**	-1.91
	(1.30)	(1.22)
$\widehat{\beta}_{3,\text{emp}}$	2.81	1.61
	(1.83)	(1.63)

Notes: The standard errors are given in parentheses, *** denotes a variable that is significant at the 1% level, ** denotes a variable that is significant at the 5% level and * denotes a variable that is significant at the 10% level.

When we apply the wild cluster bootstrap, the ATE is not significant at the thresholds examined for the DiD method. The ATE roughly equals the one obtained in Card and Krueger (1993), the standard error is however larger in this case.

When we apply the RDiD method to the data, all parameter estimates decrease in magnitude. We see a drop in the estimated treatment effect, meaning that there is even more uncertainty regarding the effect of the policy change. Based on Table 7, we suspect results of the DiD

method to be influenced by outliers. Due to the different outcomes obtained from the DiD and RDiD methods, we take a closer look at the RDiD method. As the dependent variable is continuous with all independent variables being binary, vertical outliers is the only type of outliers which can possibly be witnessed. Figure 2 shows the weights assigned to all observations by the MM-estimator used in the RDiD method, averaged over all bootstrap replications.

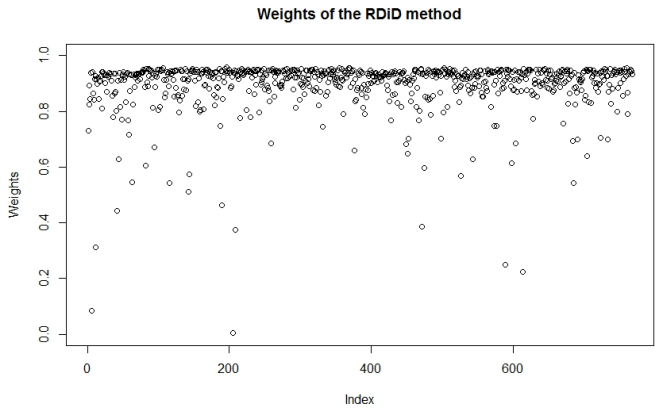

Figure 2: The observation weights obtained from the RDiD method.

Figure 2 shows that the majority of the weights lie close to one, but a good share of the weights reach values clearly lower. Some weights show to be close or equal to zero, leading to concerns regarding the original dataset. As OLS can already break down when one outlier is present, witnessing these low weights questions the appropriateness of the DiD method for this data. The distance distance plot (Rousseeuw and Van Zomeren, 1990) corresponding to the data is given in Figure 3. Robust mahalanobis distances cannot be calculated due to multicollinearity, the horizontal axis is hence given by the leverages. As the only outliers which we potentially observe are vertical outliers, the measurement given on the horizontal axis is not of much interest.

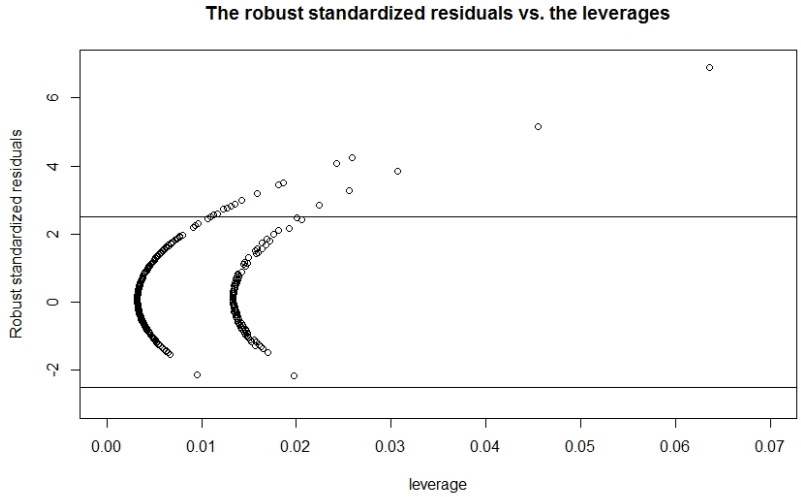

Figure 3: The distance distance plot of the RDiD method.

Figure 3 shows the presence of some vertical outliers, which are the observations above the cut off of the standardized residuals. Hence, the values of the employment variable for outlying observations deviate from the majority of this variable's values. For this particular dataset, we conclude that the DiD method falls short and that the RDiD method is the preferred one. Although the conclusion drawn based on the outcomes in Card and Krueger (1993) was already shown to be too optimistic, as discussed in Section 5.2, these results show that robust regression could have also been used in order to come to the conclusion's rejection.

6.2.2 Dinkelman (2011)

For the analysis of the data used in Dinkelman (2011) we make use of the IV type of methods. We do apply all IV type of methods to the data, but interpret all findings with caution. As already mentioned, the RIV method does not perform equally well for all data. The method of choice may therefore differ per dataset, which is why we investigate the appropriateness of the RIV method in this case as well. Before we do so, we define the models which we examine.

For the analysis we have information regarding $n = 1816$ households. Besides controls, we also have information about district fixed effects, which are the binary exogenous covariates. We index communities by j, districts by d and time periods by t, where $t = 0$ and $t = 1$ are pre and post treatment respectively. The model to be estimated is given as

$$\Delta y_{jdt+1} = \alpha_{1,\text{elec}} + \alpha_{2,\text{elec}} \Delta T_{jdt+1} + \boldsymbol{X}_{jd0}\boldsymbol{\beta}_{\text{elec}} + \lambda_{d,\text{elec}} + \big(\delta_{j,\text{elec}} + \Delta\epsilon_{jdt+1}\big), \tag{70}$$

where $\Delta x_{jdt+1} = \big(x_{jdt+1} - x_{jdt}\big)$, ϵ_{jdt} stands for the unobserved error term. Variable T_{jdt} equals one if the community has received electricity by the government at time t, and zero otherwise. The parameter vector corresponding to the controls is defined as $\boldsymbol{\beta}_{\text{elec}} = \big(\beta_{1,\text{elec}}, \ldots, \beta_{D,\text{elec}}\big)'$, if we assume that there are D districts. As we investigate two models examined in Dinkelman (2011), outcome variable y_{jdt} is given by the female and male employment rates in the separate models.

Matrix $\boldsymbol{X}_{jd0}$ contains the controls, $\lambda_{d,\text{elec}}$ and $\delta_{j,\text{elec}}$ capture community and district fixed effects respectively. The controls used are household density, the fraction of households living below a poverty line, distances to the grid, road and town, the fraction of adults that are white or Indian to proxy for local employers, the fraction of men and women with a high school certificate, the share of female-headed households and the female/male sex ratio. Treatment variable T_{jdt} is instrumented by Z_j, that is the average community land gradient.

For this dataset, we will combine the RIV method as described in Section 3.2.2 with L_1 regression as proposed in Freue et al. (2013). For this method, which we will call L_1-RIV, we define $\boldsymbol{y}_{\text{elec}}$ and $\boldsymbol{Z}_{\text{elec}}$ as the vectors stacking all outcomes and instruments over all observations respectively. Matrix $\boldsymbol{X}_{\text{elec}}$ is defined as the concatenation of all controls and the treatment, both stacked over all observations. We define $\boldsymbol{\lambda}_{\text{district}} = \big(\lambda_{1,\text{elec}}, \ldots, \lambda_{D-1,\text{elec}}\big)'$ as the vector includ-

ing district fixed effects, such that district D functions as the reference group. Furthermore, we define $\boldsymbol{X}_{\text{district}}$ as the sparse matrix, selecting the correct district fixed effect per observation, and selecting no district fixed effect if an observation belongs to district D.

The L_1-RIV method is an iterative procedure, the parameter estimates are updated according to

$$\begin{aligned}\left(\widehat{\alpha}_{2,\text{elec}}^{(q)}, \widehat{\beta}_{1,\text{elec}}^{(q)}, \ldots, \widehat{\beta}_{D,\text{elec}}^{(q)}\right)' &= \text{RIV}\left(\boldsymbol{X}_{\text{elec}}, \boldsymbol{Z}_{\text{elec}}, \boldsymbol{y}_{\text{elec}} - \boldsymbol{X}_{\text{elec}}\widehat{\boldsymbol{\lambda}}_{\text{district}}^{(q-1)}\right), \\ \widehat{\boldsymbol{\lambda}}_{\text{district}}^{(q)} &= \text{L}_1\left(\boldsymbol{X}_{\text{district}}, \boldsymbol{y}_{\text{elec}} - \boldsymbol{X}_{\text{elec}}\left(\widehat{\alpha}_{2,\text{elec}}^{(q)}, \widehat{\beta}_{1,\text{elec}}^{(q)}, \ldots, \widehat{\beta}_{D,\text{elec}}^{(q)}\right)'\right), \quad \text{for } 1 \leq q \leq Q,\end{aligned} \tag{71}$$

where we set $Q = 10$. With $\text{RIV}(\cdot)$ we denote the application of the supplied data to the RIV method as in Section 3.2.2. Note that the dependent variable is given in the latter argument, and that the intercept is not iteratively updated. The $\text{L}_1(\cdot)$ method returns parameter estimates from regressing the second argument on the first. Vector $\widehat{\boldsymbol{\lambda}}_{\text{district}}^{(0)}$ has to be initialized, we refer to Freue et al. (2013) and their Web Appendix for further details on how this is done. In the end, an estimate of $\widehat{\alpha}_{1,\text{elec}}^{(Q)}$ is obtained in a similar way as in Equations 15 and 21, now only for the model in Equation (70).

Starting off with the exploratory analysis, we determine the appropriateness of the RIV method by checking whether the data mimics the desired elliptical structure. Figure 4 shows scatterplots and spearman correlations of all variables used in the analysis, excluding the exogenous binary covariates. A derivation of how the significance levels of the correlations are determined is given in 75 in Section A. The variables from left to right are female employment rates, treatment, the controls in the same order as described earlier, and the instrument. The pairwise scatterplots of the male employment rate and all other variables barely differ from Figure 4, meaning that the conclusions drawn from Figure 4 also hold for these scatterplots. For completeness, they are given in Figure 6 in Section A.

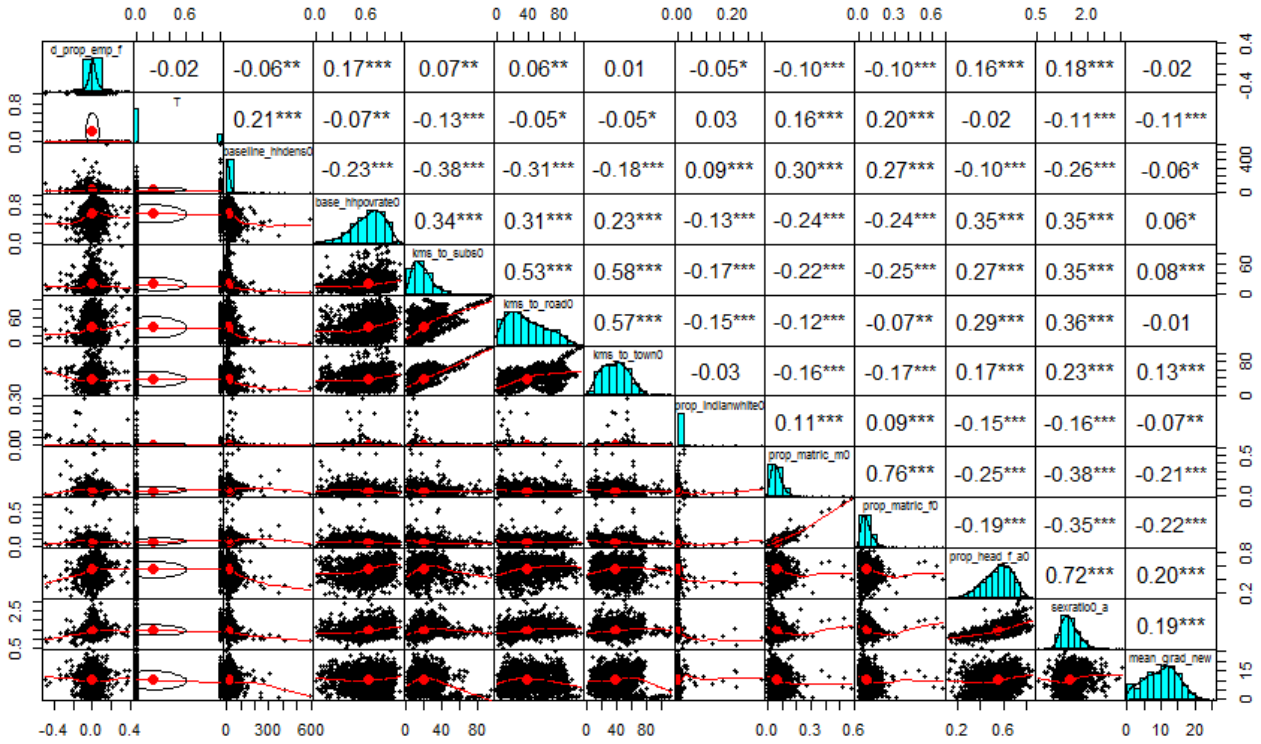

Figure 4: The pairwise scatterplots of the electrification data for the female employment rate.

Figure 4 shows elliptical type of shapes for about half of the variable pairs, the shapes of all other pairs cannot be labelled as elliptical. Hence, the data used in Dinkelman (2011) is a poor fit for the RIV method. Figure 4 does however show aberrant data points in a fair share of the plots. Some variables are relatively large in magnitude, namely household density, distances to the grid, road and town, and the average community land gradient. Most of the correlations are also significant at the 1% level, meaning that outliers affecting the results is a potential problem. Although the IV-DDC method does not necessarily outperform the IV method, it can give us additional insights about the data due to its outlier flagging abilities. As Figure 4 is difficult to interpret due to the amount of variables, we elaborate on the findings below.

We summarize the spearman correlations given in Figure 4 in Table 8. A summary for the model with the male employment rate as the dependent variable is given in Table 11 in Section A. Correlation r_{ij} is measured between variables i and j for $i \neq j$, where the same correlations are calculated as in Figure 4.

Table 8: A summary of the spearman correlations of the electrification data for the female employment rate.

		*	**	***
$\lvert r_{ij}\rvert < 0.3$	7	5	5	45
$\lvert r_{ij}\rvert \in [0.3, 0.5)$	0	0	0	11
$\lvert r_{ij}\rvert \in [0.5, 0.7]$	0	0	0	3
$\lvert r_{ij}\rvert > 0.7$	0	0	0	2
n_{cor}		78		

Notes: The first column denotes correlations which are not significant, *** denotes a correlation that is significant at the 1% level, ** denotes a correlation that is significant at the 5% level and * denotes a correlation that is significant at the 10% level. With n_{cor} we denote the amount of correlations calculated.

Table 8 shows that the majority of the correlations are weak, although most of them are significant. We focus on the bivariate pairs with correlations larger than 0.5 in absolute value, which we label as strong. The variables forming these pairs are distances to the grid, road and town, the fraction of men and women with a high school certificate, the share of female-headed households and the female/male sex ratio. A closer look at the pairwise scatterplots shows some deviating points, making it interesting for us to apply the DDC algorithm. Depending on these variables' effects on the outcome, these potential cellwise outliers may be the cause of bias in the parameter estimates.

The DDC algorithm detects 1039 cells as outliers, the variable measuring the fraction of adults that are white or Indian to proxy for local employers is dropped from the procedure, as its variation is low. The 1039 cells detected as outlying are spread over 704 observations, meaning that the DDC algorithm detects outlying cells in nearly 40% of the observations. The frequencies of outliers found in the variables are visualized in Figure 5. The order of the variables is the same as for Figure 4, the treatment variable is now however omitted. A histogram of the cellwise outliers of the electrification for the male employment rate is given in Figure 7 in Section A.

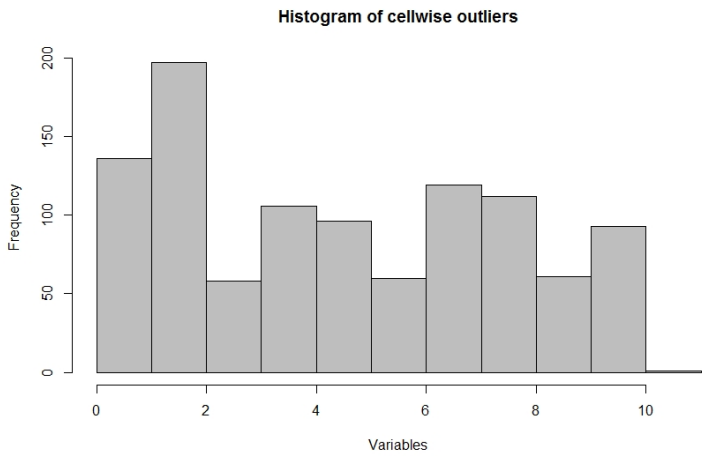

Figure 5: The histogram of the cellwise outliers of the electrification data for the female employment rate.

Figure 5 shows similar frequencies across the variables, except for the variable measuring household density. This variable is also the largest in magnitude, meaning that the cells detected as outlying could potentially form a problem.

We also attempt to measure imputation quality, although this procedure is difficult for real data. We denote the ARE and MAE of the imputations as defined in Section 4.4, these measurements are however slightly adjusted for this dataset. First of all, we set one of the cells, which are all known in this dataset, to missing in 7.5% of the observations according to the MAR mechanism. This way, we can determine how kNN imputation would perform if there would be missing values which are MAR. We split the variables for error calculation, as the appropriateness of the ARE and MAE metrics depends on the type of data.

We calculate the ARE for the variables which are large in magnitude, that is household density, distances to the grid, road and town and finally the average community land gradient. The remaining continuous variables are used for calculating the MAE, except for the fraction of adults that are white or Indian to proxy for local employers as it does not show enough deviations. The values of the MAE show no outliers across the bootstrap replications, we hence calculate it according to Equation (68). The ARE shows some outliers across the replications, we hence take the median except of the mean over all replications in order to end up with the final imputation error.

Table 9 shows the results obtained from the IV type of methods for $R = 1000$, where the outcome is given by the female employment rate. We apply the cluster bootstrap as in Cameron et al. (2008), applying a more sophisticated version of the bootstrap is superfluous as the number of groups is large enough in this dataset. For the RIV method, we set π as in Section 3.2.2 equal to 0.825, and omit bootstrap samples where multicollinearity is an issue. Lower values of π results in too much bootstrap samples suffering from multicollinearity, with $\pi = 0.825$ we

obtain estimates in 853 of the 1000 bootstrap replications. For convenience, we display the same results as those reported in the tables of interest in Dinkelman (2011).

Table 9: The results from the IV type of methods applied to the electrification data for the female employment rate.

	IV	IV-DDC	RIV	PLIV	Imputation ARE	Imputation MAE
T	0.106 (0.089)	0.082* (0.042)	0.009 (0.806)	0.204 (1.894)		
Poverty	0.032*** (0.012)	0.016** (0.008)	-1.079 (35.972)	-		
Female HH	0.033 (0.027)	0.019 (0.016)	-0.018 (0.450)	-	0.852	0.082
Sex ratio	0.031** (0.014)	0.016** (0.007)	-0.001 (0.049)	-		

Notes: The standard errors are given in parentheses, *** denotes a variable that is significant at the 1% level, ** denotes a variable that is significant at the 5% level and * denotes a variable that is significant at the 10% level.

Table 9 shows similar results as those reported in Dinkelman (2011) for the IV method, except for the ATE. The ATE was found to be just significant at the 10% level in Dinkelman (2011), when applying the clustered block bootstrap we obtain an ATE which is not significant at the 10% level. Note that R may be interpreted as too low in order to make such a claim, but results for larger values of R, which we do not include for convenience, support this claim.

When we look at the results from the IV-DDC method, we see a drop in magnitude for all parameters. When we separately generate missing values, we see that the variables concentrated around zero are imputed well, with an MAE of 0.082. The other continuous variables are however imputed rather poorly, with an ARE of 0.852. As the differences in magnitude of the parameter estimates are small, and the variables used for calculating the ARE are the largest in magnitude, we are sceptical about the performance of kNN imputation. Overall, the IV-DDC parameter estimates seem plausible, but the IV parameter estimates are more reliable in this case.

The third column of Table 9 shows the results obtained from the RIV estimator, which are non-credible. They show no similarities whatsoever when compared to the results from the IV and IV-DDC methods. When analyzing the results from all 853 replications which did not suffer from multicollinearity, the RIV method has shown to have broken down multiple times. This explains the relatively large standard errors, especially for poverty. Due to the inappropriateness of the RIV method, it is not of interest to filter out the corrupted replications, as the results which would remain would still not be credible. Although this estimator is the ideal method when it comes to robust IV estimation, this example just shows how carefully the results should be interpreted.

Finally, the ATE obtained from the PLIV method is given in the last column of Table 9. The estimated effect is larger than what is obtained from the other methods, but the associated standard error is relatively large, and the ATE is not significant at the 10% level. Analyzing the bootstrap replications revealed that the PLIV method also broke down multiple times, explaining its relatively large standard error. As our simulation study showed no superiority of the PLIV method when compared to the IV method, it is not of interest to filter out the replications in which the PLIV method broke down.

Overall, we base our conclusions on the IV and IV-DDC methods and see slight influences of outliers in the data. Above all, these outliers do not seem to distort the general patterns as already observed in Dinkelman (2011), but we cannot say this with certainty as the quality of the imputations remains questionable.

Table 10 shows the results obtained from the IV type of methods for $R = 1000$, where the outcome is now given by the male employment rate.

Table 10: The results from the IV type of methods applied to the electrification data for the male employment rate.

	IV	IV-DDC	RIV	PLIV	Imputation ARE	Imputation MAE
T	0.030 (0.080)	0.069 (0.060)	-0.260 (5.538)	0.230 (2.547)		
Poverty	0.064*** (0.016)	0.058*** (0.011)	2.393*** (49.102)	-		
Female HH	0.225*** (0.030)	0.134*** (0.024)	-0.093 (2.041)	-	0.842	0.086
Sex ratio	0.017 (0.014)	0.024** (0.010)	0.011 (0.256)	-		

Notes: The standard errors are given in parentheses, *** denotes a variable that is significant at the 1% level, ** denotes a variable that is significant at the 5% level and * denotes a variable that is significant at the 10% level.

Table 10 shows results similar to those obtained in Dinkelman (2011) for the IV method. For the IV-DDC method, the major difference lies in the parameter estimate corresponding to the ratio of households which are female headed. This measured effect of a female headed household on the change in the male employment rate is slightly more than half the magnitude when we combine IV with the DDC, while remaining significant at the 1% level. The estimated treatment effect is again not significant, meaning that the main conclusions drawn in Dinkelman (2011) based on this dataset remain valid. It is also noteworthy that the parameter estimate corresponding to the male/female ratio now is significant at the 5% level. These parameter estimates again seem plausible, but the arguments as in the description of the results of Table 9 also apply here.

The results of the RIV method show no overlap with those obtained from the other methods, which is again due to the non-elliptical structure of the data. Large standard errors are again due to the method breaking down in some of the replications, we hence regard the results obtained as non-credible.

For the PLIV method, the same conclusions can be drawn from Table 10 as is done for the results obtained from Table 10. In this simulation, the method also broke down multiple times.

We follow the same strategy as with Table 9 to conclude on our findings. Combining the IV method with the DDC algorithm leads to slightly different results, meaning that the claims made in Dinkelman (2011) hold.

7 Discussion

In this paper we have examined robustness properties of causal econometric and ML methods for contaminated data. In particular, we have investigated the IV, DiD, ARB and DML methods and proposed robust alternatives for both the IV and DiD methods. Results show that the RDiD is indeed a DiD type of method resistant to outliers. Based on our simulation study, the RIV method based on the work of Freue et al. (2013) also shows to be a robust alternative of the IV method, but it is only applicable in a limited amount of cases. Its performance heavily relies on the structure of the data, the results from the RIV method become unreliable if the data is not elliptically structured.

Besides the RIV method, we have proposed the IV-DDC method, which incorporates the work of Rousseeuw and Bossche (2018). Our simulation study however shows that this method may perform worse than the IV method, when kNN imputation is used for imputing cells flagged as outliers. Due to computational limits, we chose to make use of kNN imputation instead of a more complex imputation technique. The IV-DDC method may therefore potentially outperform the IV method, depending on the imputation technique used.

Additional to our simulation study, we have applied some of the methods investigated in this paper to two real datasets, namely the ones used in Card and Krueger (1993) and Dinkelman (2011). We applied the DiD type of methods to the data used in Card and Krueger (1993), and concluded that the data used suffers from vertical outliers. Usage of the RDiD led to shrinkage of the parameter estimates towards zero. We applied the IV type of methods to the data used in Dinkelman (2011), which required a more careful approach as the performances of these methods highly depend on the data. An exploratory analysis revealed that the data does not suit the RIV method. An application of the IV-DDC method revealed numerous cellwise outliers, but kNN imputation was also shown to lead to relatively poor imputations. Hence, applying the IV type of methods to the data used in Dinkelman (2011) emphasizes the caution which should be used when applying robust alternatives.

The main limitation of our paper lies in the simulation study. The amount of bootstrap replications and simulation runs are sufficient, but can be increased for more accurate results. Besides, we have not investigated high-dimensional problems, for which the causal ML methods are especially suited. Simulation studies can always be extended, but we mainly encourage the investigation of high-dimensional problems.

As outliers and missing data are two phenomena scarcely researched in the causal inference setting, there are numerous interesting directions for further research. First of all, using different imputation techniques with the IV-DDC method can potentially lead to better performances compared to the IV method, see Osman et al. (2018) for a survey of frequently used techniques. Also, the cellwise outliers generated in our simulation study were extreme. Examining all methods when outliers are less extreme may reveal some interesting patterns. Finally, developing robust alternatives of the causal ML methods, or focusing on robustification in general is an interesting topic for researchers in the field of ML. Causal ML methods are extensively developed at the moment, we can hence imagine that developing robust alternatives is not yet of interest as there are numerous unexplored ways for coming up with a new type of method.

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Athey, S., Imbens, G. W., and Wager, S. (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623.

Belloni, A., Chernozhukov, V., et al. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, 119(1):249–275.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association*, 90(430):443–450.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge university press.

Card, D. and Krueger, A. B. (1993). Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania. Technical report, National Bureau of Economic Research.

Card, D. and Krueger, A. B. (2000). Minimum wages and employment: a case study of the fast-food industry in new jersey and pennsylvania: reply. *American Economic Review*, 90(5):1397–1420.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Dinkelman, T. (2011). The effects of rural electrification on employment: New evidence from south africa. *American Economic Review*, 101(7):3078–3108.

Donald, S. G. and Lang, K. (2007). Inference with difference-in-differences and other panel data. *The review of Economics and Statistics*, 89(2):221–233.

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. *A festschrift for Erich L. Lehmann*, 157184.

Efron, B. (1979). Computers and the theory of statistics: thinking the unthinkable. *SIAM review*, 21(4):460–480.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American statistical association*, 78(382):316–331.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer.

Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, 89(426):463–475.

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560.

Feng, J., Xu, H., Mannor, S., and Yan, S. (2014). Robust logistic regression and classification. In *Advances in neural information processing systems*, pages 253–261.

Freue, G. V. C., Ortiz-Molina, H., and Zamar, R. H. (2013). A natural robustification of the ordinary instrumental variables estimator. *Biometrics*, 69(3):641–650.

Frölich, M. (2008). Parametric and nonparametric regression in the presence of endogenous control variables. *International Statistical Review*, 76(2):214–227.

Han, J. S., Houde, J.-F., van Benthem, A. A., and Abito, J. M. (2018). Difference-in-differences estimation in the presence of outliers: New evidence on the cost savings of divestiture. Technical report, Working Paper.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer.

Huber, P. J. et al. (1973). Robust regression: asymptotics, conjectures and monte carlo. *The Annals of Statistics*, 1(5):799–821.

Little, R. J. and Rubin, D. B. (2002). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Mammen, E. (2012). *When does bootstrap work?: asymptotic results and simulations*, volume 77. Springer Science & Business Media.

Neumark, D. and Wascher, W. (2000). Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania: Comment. *American Economic Review*, 90(5):1362–1396.

Neyman, J. (1959). Optimal asymptotic tests of composite hypotheses. *Probability and statsitics*, pages 213–234.

Neyman, J. (1979). C (α) tests and their use. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 1–21.

Osman, M. S., Abu-Mahfouz, A. M., and Page, P. R. (2018). A survey on data imputation techniques: Water distribution system as a use case. *IEEE Access*, 6:63279–63291.

Pison, G., Van Aelst, S., and Willems, G. (2002). Small sample corrections for lts and mcd. *Metrika*, 55(1-2):111–123.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954.

Rousseeuw, P. and Yohai, V. (1984). Robust regression by means of s-estimators. In *Robust and nonlinear time series analysis*, pages 256–272. Springer.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8(283-297):37.

Rousseeuw, P. J. and Bossche, W. V. D. (2018). Detecting deviating data cells. *Technometrics*, 60(2):135–145.

Rousseeuw, P. J. and Christmann, A. (2003). Robustness against separation and outliers in logistic regression. *Computational Statistics & Data Analysis*, 43(3):315–332.

Rousseeuw, P. J. and Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust regression and outlier detection*, volume 589. John wiley & sons.

Rousseeuw, P. J. and Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical association*, 85(411):633–639.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Templ, M., Kowarik, A., and Filzmoser, P. (2011). Iterative stepwise regression imputation using standard and robust methods. *Computational Statistics & Data Analysis*, 55(10):2793–2806.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525.

Von Hippel, P. T. (2007). 4. regression with missing ys: An improved strategy for analyzing multiply imputed data. *Sociological Methodology*, 37(1):83–117.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Wright, P. G. (1928). *Tariff on animal and vegetable oils*. Macmillan Company, New York.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, pages 642–656.

A Appendix

The proof of $\boldsymbol{P_Z}$ being idempotent is given as

$$\begin{aligned}\boldsymbol{P_Z^2} &= \boldsymbol{Z}\big(\boldsymbol{Z}'\boldsymbol{Z}\big)^{-1}\boldsymbol{Z}'\boldsymbol{Z}\big(\boldsymbol{Z}'\boldsymbol{Z}\big)^{-1}\boldsymbol{Z}' \\ &= \boldsymbol{Z}\boldsymbol{I}_{k+1}\big(\boldsymbol{Z}'\boldsymbol{Z}\big)^{-1}\boldsymbol{Z}' \\ &= \boldsymbol{Z}\big(\boldsymbol{Z}'\boldsymbol{Z}\big)^{-1}\boldsymbol{Z}' \\ &= \boldsymbol{P_Z}.\end{aligned} \tag{72}$$

The proof of $\boldsymbol{P_Z}$ being symmetric is given as

$$\begin{aligned}\boldsymbol{P}_{\boldsymbol{Z}}' &= \big(\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\big)' \\ &= \boldsymbol{Z}\big((\boldsymbol{Z}'\boldsymbol{Z})^{-1}\big)'\boldsymbol{Z}' \\ &= \boldsymbol{Z}\big((\boldsymbol{Z}'\boldsymbol{Z})'\big)^{-1}\boldsymbol{Z}' \\ &= \boldsymbol{Z}\big(\boldsymbol{Z}'\boldsymbol{Z}\big)^{-1}\boldsymbol{Z}' \\ &= \boldsymbol{P_Z}.\end{aligned} \tag{73}$$

The full derivation of Equation (20) is given as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\mathrm{IV}} &= \big(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}}\big)^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y} \\ &= \big(\boldsymbol{X}'\boldsymbol{P}_Z'\boldsymbol{P_Z}\boldsymbol{X}\big)^{-1}\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}'\boldsymbol{y} \\ &= \big(\boldsymbol{X}'\boldsymbol{P_Z}\boldsymbol{X}\big)^{-1}\boldsymbol{X}'\boldsymbol{P_Z}\boldsymbol{y} \\ &= \big(\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}\big)^{-1}\boldsymbol{X}'\boldsymbol{Z}\big(\boldsymbol{Z}'\boldsymbol{Z}\big)^{-1}\boldsymbol{Z}'\boldsymbol{y} \\ &= \big(n\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{XZ}}\ \frac{1}{n}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{ZZ}}^{-1}\ n\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{ZX}}\big)^{-1}\ n\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{XZ}}\ \frac{1}{n}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{ZZ}}^{-1}\ n\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Zy}} \\ &= \big(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{XZ}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{ZZ}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{ZX}}\big)^{-1}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{XZ}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{ZZ}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Zy}}.\end{aligned} \tag{74}$$

The pairwise scatterplots of the electrification data for the male employment rate are given in Figure 6.

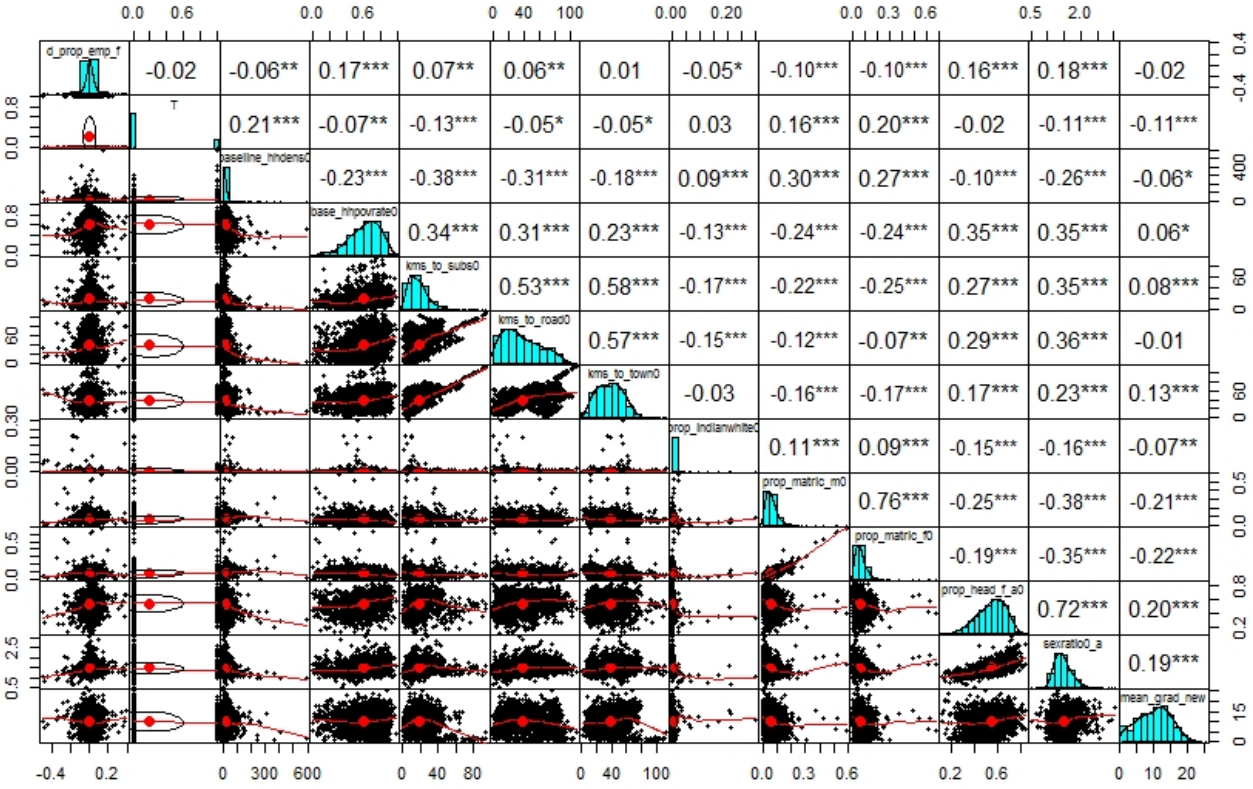

Figure 6: The pairwise scatterplots of the electrification data for the male employment rate.

A summary of the spearman correlations of the electrification data for the male employment rate is given in Table 11.

Table 11: A summary of the spearman correlations of the electrification data for the female employment rate.

		*	**	***
$\lvert r_{ij}\rvert < 0.3$	5	5	3	47
$\lvert r_{ij}\rvert \in [0.3, 0.5)$	0	0	0	13
$\lvert r_{ij}\rvert \in [0.5, 0.7]$	0	0	0	3
$\lvert r_{ij}\rvert > 0.7$	0	0	0	2
n_{cor}		78		

Notes: The first column denotes correlations which are not significant, *** denotes a correlation that is significant at the 1% level, ** denotes a correlation that is significant at the 5% level and * denotes a correlation that is significant at the 10% level. With n_{cor} we denote the amount of correlations calculated.

The histogram of the cellwise outliers of the electrification data for the female employment rate are given in Figure 7.

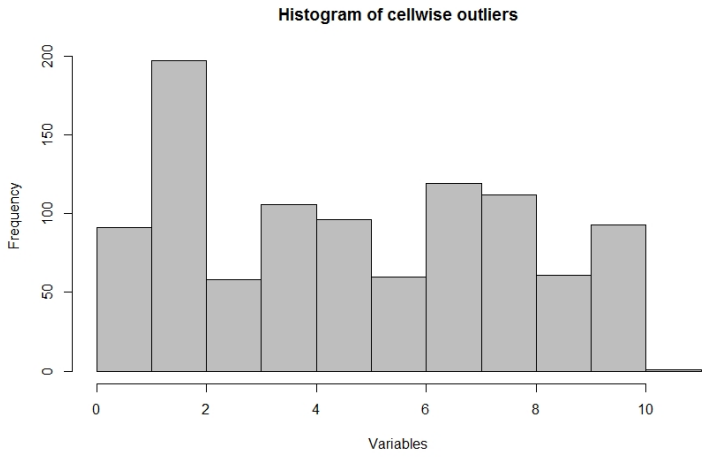

Figure 7: The histogram of the cellwise outliers of the electrification data for the male employment rate.

Under the null hypothesis of zero population correlation, the test statistic used for determining sample correlation significance is given as

$$t_{ij} = \frac{r_{ij}}{\sqrt{1 - r_{ij}^2}} \cdot \sqrt{n-2}. \tag{75}$$

When we conduct a two-tailed test, the null hypothesis is in turn rejected if $\left|t_{ij}\right| > t_{\alpha/2,n-2}$, where α is the significance level.

The results from the DiD type of methods for $\epsilon_{\text{cont}} = 20\%$, where outliers are caused due to bad leverage points are given in Table 12.

Table 12: The results of the DiD type of methods for $\epsilon_{\text{cont}} = 20\%$ and outliers due to bad leverage points.

	RMSE	Bias	Coverage	ATE		Prediction	Outlier Detection	Imputation ARE	Imputation MAE
DiD	2.221	0.044	47.250%	4.679	(2.312)	89.482	-	2.383	0.559
RDiD	**0.210**	0.044	**93.375%**	**4.148**	(0.297)	**1.068**	0.999		

Notes: The standard errors are given in parentheses.

The results from the IV type of methods for $\epsilon_{\text{cont}} = 20\%$, where outliers are caused due to bad leverage points are given in Table 13.

Table 13: The results of the IV type of methods for $\epsilon_{\text{cont}} = 20\%$ and outliers due to bad leverage points.

	RMSE	Bias	Coverage	ATE		Prediction	Outlier Detection	Imputation ARE	Imputation MAE
IV	17.916	-1.251	39.400%	-2.546	(34.954)	1769.586	-	3.522	0.664
RIV	**0.675**	**0.030**	**95.400%**	**3.844**	(1.210)	**159.924**	0.805		
IV-DDC	14.039	-1.291	37.600%	-2.701	(27.294)	1157.876	**0.810**		
PLIV	20.333	-0.680	92.000%	3.092	(20.424)	-	-		

Notes: The standard errors are given in parentheses.

The results from the ML type of methods for $\epsilon_{\text{cont}} = 20\%$, where outliers are caused due to bad leverage points are given in Table 14.

Table 14: The results of the ML type of methods for $\epsilon_{\text{cont}} = 20\%$ and outliers due to bad leverage points.

	RMSE	Bias	Coverage	ATE		Imputation ARE	Imputation MAE
ARB	1.276	0.210	**97.000%**	3.981	(1.265)	2.751	0.515
PLR	**1.161**	**0.137**	92.000%	**3.908**	(1.159)		

Notes: The standard errors are given in parentheses.

The results from the DiD type of methods for $\epsilon_{\text{cont}} = 20\%$, where outliers are caused due to vertical outliers are given in Table 15.

Table 15: The results of the DiD type of methods for $\epsilon_{\text{cont}} = 20\%$ and outliers due to vertical outliers.

	RMSE	Bias	Coverage	ATE		Prediction	Outlier Detection	Imputation ARE	Imputation MAE
DiD	9.207	2.482	83.250%	4.722	(9.048)	475.136	-	2.196	0.551
RDiD	**0.214**	**0.046**	**93.750%**	**4.115**	(0.299)	**1.545**	1.000		

Notes: The standard errors are given in parentheses.

The results from the IV type of methods for $\epsilon_{\text{cont}} = 20\%$, where outliers are caused due to vertical outliers are given in Table 16.

Table 16: The results of the IV type of methods for $\epsilon_{\text{cont}} = 20\%$ and outliers due to vertical outliers.

	RMSE	Bias	Coverage	ATE	Prediction	Outlier Detection	Imputation ARE	Imputation MAE
IV	11.071	3.851	79.000%	3.457 (11.357)	806.375	-	3.289	0.661
RIV	0.764	0.067	**94.200%**	**3.736** (1.208)	**130.440**	**0.993**		
IV-DDC	**0.741**	**0.028**	93.600%	3.812 (1.325)	153.291	0.969		
PLIV	11.315	-0.458	100.000%	3.314 (21.156)	-	-		

Notes: The standard errors are given in parentheses.

The results from the ML type of methods for $\epsilon_{\text{cont}} = 20\%$, where outliers are caused due to vertical outliers are given in Table 17.

Table 17: The results of the ML type of methods for $\epsilon_{\text{cont}} = 20\%$ and outliers due to vertical outliers.

	RMSE	Bias	Coverage	ATE	Imputation ARE	Imputation MAE
ARB	3.967	**-0.032**	96.000%	**3.739** (3.987)	2.761	0.495
PLR	**3.902**	-0.047	96.000%	3.724 (3.921)		

Notes: The standard errors are given in parentheses.