Erasmus University Rotterdam
Business Analytics and Quantitative Marketing
FEM61008

Evaluating Treatment-effect Methods for Contaminated Data: an Application of Influence- and Change-of-Variance Functions

Research question

"What are good estimators to use for contaminated, multivariate data in a randomised experiment, based on the influence- and change-of-variance function?"

Author:
Daphne Moors (448747dm)

Supervisor and second assessor:
Mikhail Zhelonkin & Nick Koning

October 25, 2023

Abstract

Treatment effect estimation is a popular field of study, but the frequently used classical methods like OLS and mean-based estimators break down in the presence of outliers. Outliers can be difficult to detect and remove, therefore caution is warranted when using those methods. Furthermore, a theoretical understanding of the effects of infinitesimal levels of contamination on treatment effect estimates value and, most specifically, variance is not sufficiently studied. This thesis derived expressions for the influence function (IF) and change-of-variance (CVF) function for three different treatment-effect estimators: the difference-in-means estimator, the regression-adjusted estimator and the difference-in-intercepts estimator. Additionally, this thesis compares the classical estimators with (more) robust alternatives in a simulation study and a real-life data application. It is shown that classical estimators are biased for infinitesimal levels of contamination, and that robust alternatives are promising.

The content of this thesis is the sole responsibility of the author and does not reflect the view of the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

Contents

1 Introduction

Everywhere we look around us, whether it be in business, personal or governmental settings, we are interested in knowing about the results of "doing something rather than doing something else". This can be as small as wondering about the health benefits of cycling to work every day instead of taking the car or as big as finding the effect of implementing new environmental European Union legislation on CO_2-emissions. The effect(s) of a treatment or change on a response or outcome variable is called the *treatment effect*. Because it is such a generic concept, interest for the effect of a treatment on a response variable is found in many areas of research as well, including medicine (Hollingsworth et al., 2006; Dechartres et al., 2013; and Austin and Stuart, 2015), the social and economic sciences (Angrist, 2004; Sampson et al., 2006; and King et al., 2007) and policy making (Johansson and Palme, 2002; Shambaugh, 2004; and Criscuolo et al., 2012).

Though treatment effects have many interested, its estimation can sometimes still cause trouble. The fundamental problem is that of the unobservable *counterfactual* (i.e. what would have been the individual's outcome in case they had (not) been given the treatment, instead). A naive comparison between observations with and without the treatment, leaves the door open to selection bias. The 'gold-standard' to solve the selection bias problem is to opt for randomised controlled trials (RCT). In the field of economics and econometrics, (large-scale) experiments are often very costly and researchers have to work with observational data instead, and then apply statistical methods (e.g. matching) to overcome this problem. However, in other fields of research, like medicine, RCTs are more the norm. This paper focuses on applications for RCTs and methods for working with observational data are beyond the scope of this paper.

Even though the problem of selection bias can be solved by using RCTs, another problem often occurs when moving from simulations to real data: outliers. Outliers are deviating observations that threaten the (distributional) robustness: "the shape of the true underlying distribution deviates slightly from the assumed model" (Huber and Ronchetti, 2009, Chapter 1). In real-life data sets, outliers are nothing out of the ordinary (Rousseeuw and Leroy, 1987; and Zaman et al., 2001) and much research is performed to tackle this issue (e.g. Hampel et al., 1986; Rousseeuw and Leroy, 1987; Barnett et al., 1994; and Rocke and Woodruff, 1996). Overviews of possible outliers, detection methods and applications can be found in the research articles by Hodge and Austin (2004); Rousseeuw and Hubert (2018); and Grentzelos et al. (2021).

An intuitive illustration emphasising the critical importance for robust methods is one about governmental subsidies to low-income households: In general, the target group has a lower level of education (if any at all), but some individuals eligible for the financial assistance are highly educated entrepreneurs (i.e. the outliers) with not yet profitable start-ups. For the majority of the group it is likely that the aid is used for paying off debts and making ends meet. However, let us now assume that the entrepreneurs can attain large returns on their financial assistance because of successful investments. Evaluating the assistance' success with classic instruments, the few entrepreneur's high returns of investment will draw the treatment estimate for the entire group towards them and the researcher is left with a (positively) biased conclusion. Overcoming problems like such is a challenge that is met by the field of robust statistics and is critical to many studies, as classical approaches like OLS and popular estimators such as the sample mean can break down in the presence of just a single outlier (Rousseeuw and Wagner, 1994; and Rousseeuw and Leroy, 1987, Chapter 2 & Baharudin et al., 2012, respectively). One example of a robust estimator is the MM-estimator (Yohai, 1987), showing both high breakdown properties and being efficient.

The general problem regarding outliers is that the observed data does not follow the assumed underlying distribution perfectly and that, in reality, it is almost always impossible to know the true data generating process. As a result, wrongful assumptions leading to the usage of an improper model can increase the bias and variance of an estimator significantly. In general, but also in treatment effect studies, prevention hereof is critical for a proper analysis. This thesis takes a theoretical approach using influence functions (IF) and change-of-variance functions (CVF), which are expressions to evaluate an estimator's robustness to infinitesimal levels of contamination. More specifically, this thesis evaluates the effect of just a few outliers in a sample on the value and variance of several M-type (Huber, 1964) treatment-effect estimators. To do so, the Tukey-Huber contamination model (Tukey, 1962; Huber, 1964) is used to introduce this distributional contamination. If an estimator's IF and CVF approach infinity, the estimator's value or variance is not robust to outliers, respectively.

In treatment effect analysis, there are generally two estimating methods: The first one is the simple, classical difference-in-means estimator (unadjusted) and the second is a regression-adjusted estimator, where additionally covariates are included. The benefit of using the former is that it is unbiased, whereas the latter is consistent, but biased in smaller samples. On the other hand, the regression-adjusted estimator can give more precise estimates due to a lower estimator variance (Lin, 2013). In the context of outliers, it is therefore interesting to explore the differences of the effect of outliers between the classical difference-in-means estimator and regression adjusted techniques. For example, if the regression-adjusted estimator's CVF is unbounded, the estimator's variance can inflate, thus making regression-adjusted useless.

As an addition to the scientific literature in this field, this thesis explores the IF and CVF of three different treatment effect estimators; one unadjusted and two regression-adjusted estimators. Firstly, the classical average treatment effect estimator (the difference-in-means estimator) $\bar{Y}_1 - \bar{Y}_0$ is explored. Next, the first regression-adjusted estimator (Lin, 2013) is the estimated coefficient of the treatment indicator variable in the regression of the response variable on the binary treatment indicator variable T, covariates X and an interaction between the two (now with demeaned covariates). The second regression-adjusted estimator is a two-stage estimator (Liu and Yang, 2020; Lei and Ding, 2020), where the response variable Y is regressed on (demeaned) covariates X for the treatment and control group separately, after which the difference between the regression constants is computed. Since all three estimators are M-type estimators, the IFs and CVFs are derived following the approach and notations of Zhelonkin (2013). Additionally, robust alternatives are investigated and compared to the classical estimators in a simulation study and a data application. For the unadjusted estimator, the robust alternatives evaluated are the difference-in-trimmed-means and the difference-in-medians estimators and the MM- and the Krasker-Welsch-estimator (KW-estimator) are discussed for the regression-adjusted estimators.

As a line of focus, this thesis investigates the following main question:

"What are good estimators to use for contaminated, multivariate data in a randomised experiment, based on the influence- and change-of-variance function?"

This main question can be further divided into more specific sub-questions:

- *What expressions represent the influence- and change-of-variance functions for different M-type treatment-effect estimators?*
- *Do different types of outliers affect the influence- and change-of-variance function differently? If so, what are the differences?*
- *What are more robust alternatives to the naive treatment-effect estimators and how do they change both functions?*
- *How do the theoretical expressions relate to the practical applications and results?*

A general conclusion is that it is advisable to use robust alternatives to the classical estimators. The median has bounded IF and CVF and the MM-estimator is also robust to outliers (good leverage points excluded). Though both options are promising, it is critical to be mindful about their methods and the way outliers affect the estimates. The median estimate is strongly affected by the distribution (density) around the "clean" median in the presence of outliers and can still provide biased results, especially for smaller treatment effects. The MM-estimator gives less biased results, on average, and has a lower variance, but it can underestimate the variance in the presence of good leverage points. Moreover, regression-adjusted estimators do not, per definition, lose their advantage of increased estimate precision (compared to the unadjusted estimator) when outliers are present. This strongly depends on the type and strength of the contamination.

The remainder of this paper is structured as follows: Section 2 continues discussing the developments of robust statistics and presents the literature introducing the different types of treatment effect estimators. Section 3 provides theory on specific topics regarding outliers and treatment effects, including details about influence- and change-of-variance functions and the Neyman-Rubin causal model. Next, Section 4 summarises the IF- and CVF-derivations and comments on their robustness. After this, Section 5 describes the experimental settings for the simulation study and the real-data analysis, alongside a description of this data set. The methodology for the analysis is provided in Section 6. Thereafter, the study results are presented in Section 7 and a discussion and conclusion is shared in Section 8.

2 Literature Review

This section provides an overview of the literature in the field of treatment effect evaluation and robust statistics. Firstly, the central principle of treatment effects and randomised studies is discussed in Section 2.1. Next, Section 2.2 provides an overview of different estimators that can be used for treatment effect analysis. Lastly, Section 2.3 gives a brief overview of some robust statistics' developments, including a section discussing influence- and change-of-variance functions.

2.1 Treatment effect & causal inference

Treatment effect and causal inference are central concepts in many areas of statistics and data analysis, including medicine, economics, and the social sciences (see the aforementioned examples in Section 1. The goal of treatment effect analysis is to determine the causal effect of a treatment or intervention on an outcome variable, by comparing the outcomes of treated and untreated individuals. The true treatment effect is described as: $\tau = E[Y_{i1} - Y_{i0}]$, where Y_1 represents the outcome variables for the treatment group and Y_0 represents the control group. However, for individual i, observations Y_{i1} and Y_{i0} can never both be observed. This is known as the *counterfactual* problem. It is therefore crucial to be aware of the principles laid out in the well-known Neyman-Rubin causal model, as described by Holland (1986). This is an approach based on a potential outcomes framework first introduced by Neyman (1923) and work by Rubin (1974, 1977, 1978, 1980). This thesis also operates within this framework.

Naively comparing results between treatment and control groups can lead to biased conclusions. The self-selection into treatment (or control) causes a disbalance between (unobserved) covariates between the groups. Besides the treatment, other covariates also have predictive power towards the outcome variable, and hence this introduces selection bias. In observational studies, there is much research in the field of covariates balancing (e.g. Imai and Ratkovic, 2014) to overcome this problem. However, the 'gold standard' to estimate causal effects within treatment effect analysis are RCTs, where individuals are randomly assigned to either the treatment or control group. RCTs minimise (on average) selection bias and ensure that any differences in outcomes between the groups are due to the treatment and not other factors, again, on average. This statement strongly relates to the *independence assumption* $E(Y_t) = E(Y_t|T = t)$ for $t = 0, 1$ (control and treatment group, respectively) as laid out by Holland (1986). In turn, this allows for a direct comparison between the treatment and control group. An extensive review of the use (and misuse) of RCTs is provided by Deaton and Cartwright (2018). Though an RCT is not always feasible or ethical in all cases, there are still plenty of studies where this type of experiment is possible, such as Naci and Ioannidis (2015); Berkhemer et al. (2015); Yusuf et al. (2016); and Stolberg et al. (2018) in the field of medicine. Applications like these are also the focus of this thesis.

2.2 Different types of treatment effect estimators

This section summarises relevant developments in the field of treatment effect analysis. Specifically, it describes the differences between different estimation models in an RCT setting.

2.2.1 One-stage estimators

Treatment-effect estimation has been a field of interest ever since the introduction of the *intention-to-treat* (ITT) estimator (estimating the effect of (random) assignment to treatment) by Neyman (1923). This estimator computes the average of the outcome variables in the treatment group and subtracts the average of the control group:

$$\hat{\tau}_{unadj} = \bar{Y}_1 - \bar{Y}_0$$

In later literature, this estimator is referred to as the *unadjusted* estimator. It is shown to be an unbiased estimator for τ with variance $\frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\tau^2}{N}$ (Neyman, 1923), where S_1^2, S_0^2, and S_τ^2 are the finite-population variances of Y_{i1} and Y_{i0} and τ_i, respectively. Since S_τ^2 cannot be estimated without further assumptions, it is often removed and standard error estimation is considered conservative (Lei and Ding, 2020). Additional benefits of this estimator are that it does not requires any assumptions on model specification or statistical distributions, and the treatment effect can be either homo- or heterogeneous. It only requires the existence of (hypothetical) counterfactual observations. Deaton and Cartwright (2018) summarise the discussion regarding the validity of this assumption.

This unadjusted estimator, though unbiased, does not include any pre-treatment covariates, which when included can lower the estimator's variance if they show predictive power (Fisher, 1932). In case there is no predictive power, this estimator is asymptotically efficient. Some early works using covariates are (Fisher, 1936; and Kempthorne, 1952), where the treatment effect is commonly estimated by the coefficient of the treatment indicator in the OLS-fit of the outcome variable on the treatment indicator variable and the covariates. Estimators like these are named *regression-adjusted* estimators.

$$y = \hat{\mu} + \hat{\tau}_{RA} T + X\hat{\beta} + \hat{\varepsilon}$$

Including those additional covariates into the analysis, leads to an at least even as efficient estimator compared to the unadjusted estimator. When doing so, a constant (homogeneous) treatment effect estimator is implicitly assumed. However, in case the treatment effect are heterogeneous and the treatment and control group are unequal in size, an influential critique by Freedman (2008) shows that the estimator can be even *less* efficient than the unadjusted estimator and the variance estimate can even be inconsistent for the true variance under randomised treatment effect settings.

Reacting to Freedman (2008), Lin (2013) proposed a simple solution, introducing another estimator: a consistent and an at least as efficient estimator is the treatment effect indicator's coefficient from the OLS fit of the outcome variable on the treatment indicator, the covariates *and* an interaction term between the treatment indicator and the demeaned covariates:

$$y = \hat{\mu} + \hat{\tau}_{Lin} T + X\hat{\beta} + T(X - \bar{X})\hat{\gamma} + \hat{\varepsilon}$$

In the respective paper the estimator's name is the *interaction* estimator, but in general literature, including this thesis, this is also called a regression-adjusted estimator. Numerically, this is similar to choosing two different treatment indicator coefficients for both groups (i.e. heterogeneity). The interaction design should always be preferred unless one of the following two statements hold: the design in perfectly balanced (i.e. $N_{treatment} = N_{control}$) and there is no heterogeneity in the estimated coefficients (Negi and Wooldridge, 2021). If any of these hold true, there is no efficiency gain in using the interaction design over the regression-adjustment design without the intereaction term and the latter is even preferred to preserve degrees of freedom. Additionally, Lin (2013) shows that the Huber-White sandwich standard error (White (1980a, 1980b)) proposed a consistent covariance matrix estimator for OLS, named the Huber-White because the estimator is the sample analog of Huber's (1967) formula for asymptotic variance of the maximum likelihood estimator when the model is incorrect) estimate is consistent or asymptotically conservative. Though critiqued by Freedman (2006), Huber-White standard errors are still relevant here, as estimates are still consistent even for incorrect regression models. It should also be noted that both Freedman (2008) and Lin (2013) operate under the finite population paradigm where all population units are observed in the sample. This implies that uncertainty in the estimators is because of assignment into treatment and control and not due to sampling for a population (Negi and Wooldridge, 2021).

2.2.2 Two-stage estimators

Interestingly, Lin's estimator numerically equals a two-stage estimator, where the estimator is the difference between two (i.e. treatment and control) OLS-regression intercepts, as is performed by Liu and Yang (2020) and Lei and Ding (2020). In the first step, estimates for the coefficients' and the intercept's coefficients are obtained separately, for the treatment and control group.

$$y = \hat{\mu} + (X - \bar{X})\hat{\beta} + \hat{\varepsilon}, \text{ for both the treatment and control group separately.}$$

Then, the treatment effect is estimated by subtracting the intercept estimate of the control group of that of the treatment group:

$$\hat{\tau} = \hat{\mu}_1 - \hat{\mu}_0$$

It is implicitly assumed that the population treatment effect is a constant difference between the population treatment and control groups (see paragraph 4.4 in Holland, 1986). This two-stage estimator is also biased in finite samples, but Lei and Ding (2020) do suggest a bias-adjusted estimator based on the leverage score (hat-matrix). They also include a bias formula that extends the bias formula in Lin (2013) to the multivariate case.

2.3 Robust statistics

Robust statistics is a field of statistical theory that focuses on the development of methods that are less sensitive to outlying observations and deviations from the assumed distribution. The need for robust statistical methods arises from the fact that classical statistical methods, such as maximum likelihood estimation (MLE) (Hennig, 2004) and OLS (Rousseeuw and Leroy, 1987, Chapter 2), are often sensitive to outliers, which can have a significant impact on the estimation results. Moreover, in the field of treatment effect estimation, the term *robustness* is also applied when describing the estimator's robustness to model misspecification. Just like robustness to outliers, this is also an important aspect of unbiased estimates and a popular concept to research in the field of treatment effect. In short, *double robustness* is a widely published concept in (non-RCT) treatment effect studies, e.g. Kang and Schafer, 2007; Li et al., 2016; and Kurz, 2022, and entails that for estimations that consist of two models (i.e. outcome regression and propensity score estimation), the estimation is robust to misspecification of one of these models. Though the topics are not complete unrelated, robustness to model misspecification lies beyond the scope of this thesis.

In the field of robustness to outliers, treatment effect studies are represented in the literature, including matching (Canavire-Bacarreza et al., 2021) and RCTs (Deaton and

Cartwright, 2018) studies, but the number of papers overall is scarce. Canavire-Bacarreza et al. (2021) explore that bad leverage points give rise to bias and that good leverage points can break the common support condition and distort covariance balance between the two groups. In turn, Deaton and Cartwright (2018) show that the estimation of the treatment effect can appear bimodal, depending on the location of the outlier(s), be it in the control or treatment group, and argue that inference on means can be difficult. In any case, both papers show the importance of being aware of the presence of outliers and using appropriate methods to resolve the problem of biased estimators.

2.3.1 Influence function & Change-of-Variance function

The influence function is a way to formally express the influence of an outlying observation in the sample on the estimator's value. If the influence function is bounded, the estimator is (bias-)robust to infinitesimal levels of contamination. The first introduction of the influence function is by Hampel (1968, 1974), after which many extensive books are written about robust statistics and the influence function specifically (Hampel et al., 1986; Rousseeuw and Leroy, 1987; and Huber and Ronchetti, 2009). Similarly, the effect of outliers of an estimator's variance is interesting to analyse. This notion is also firstly introduced by Hampel (1968), later referred to as the change-of-variance function. Rousseeuw (1981a, 1981b) provided an important foundation for this concept. Both functions evaluate the effect of infinitesimal levels of contamination and therefore provide information about robustness against a single outlying observation (in practice). Therefore, conclusions about higher levels of contamination cannot be directly drawn from influence or change-of-variance functions.

A downside of the influence function is that it is unique for every estimator and that its derivation gets more complex for more complex estimators. Within the field of M-estimators, Zhelonkin (2013) provided a detailed derivation for both the IF and CVF. Overall, the change-of-variance function has received much less attention in the literature. On the one hand, this can be due to its derivational complexity. On the other hand, it can also be the natural results of being "the second step" in robustness analysis, whereas finding a consistent estimator under the influence of outliers is likely to be the first step. Furthermore, robustness evaluations of treatment effect estimation methods using IF and CVF are underpublished, let alone for analysing the effect of different types of outliers.

2.3.2 Robust statistics for linear regression

One of the earliest contributions to robust statistics is done by Huber (1964), introducing the so-called *M-estimator* in the field of regression analysis. Huber argued that the least-squares loss function is not robust to outliers and suggests the use of more robust loss functions, instead. The general class of M-estimators for $\hat{\beta}$ is defined as

$$\sum_{i=1}^{N} \phi(y_i, x_i, \hat{\beta}) = 0,$$

for some non-constant function ϕ. The least squares estimator is defined by equation $\sum_{i=1}^{N}(y_i - x_i\hat{\beta})x^t = 0$, and it is therefore evident that it is part of the M-type estimators with $\phi(y, x, \beta) = (y - x\beta)x^t$. However, these estimators are only robust to outliers in the response variable, thus only improving robustness marginally. Furthermore, an important extension of the introduced M-estimators by Huber, is the contribution of Tukey (1977), introducing the Tukey loss function, also known as the Tukey biweight or bisquare loss function. This loss function gained much popularity, as the loss function is truncated at a threshold (i.e. bounded, and thus robust) using to a tunable parameter c and is also robust to leverage points.

Krasker (1980) and Krasker and Welsch (1982) obtained the optimal bounded-influence estimators (the Hampel-Krasker (HK) estimator and the Krasker-Welsch (KW) estimator, respecively). These estimators are called "optimal" because, next to their influence curves, also their change-of-variance functions are bounded, as derived by Ronchetti and Rousseeuw (1985). However, it must be noted that Maronna et al. (1979) showed that Generalised M-estimators (GM-estimators) break down when the number of covariates $p > 1$ (the maximum breakdown point is $\frac{1}{p+1}$), which includes the KW-estimator mentioned before. Therefore, it is important to note that bounded IF and CVF do not imply a high breakdown point (i.e. robustness to higher levels of contamination). The KW-estimator is still considered optimal robust, but is not much applied in current research. As Flavin (1999) phrased it: "it has not passed the basic 'market test'".

Rousseeuw (1984) approached the problem of robustness through a different lens: he critiqued the "least squares" popularity and robustified the estimator through the "sum" instead of the "squares", introducing the least median of squares (LSM). The LMS-estimator reaches a 50% breakdown point, though it has a very low efficiency. In the same paper, he also introduces the *least trimmed squares* (LTS) estimator, which also has a 50% breakdown point but similar efficiency as an M-estimator. A downside of this alternative is its exponentially growing computation time for larger data sets. However, in Rousseeuw and van Driessen (2005) the FAST-LTS algorithm is introduced to solve the issue. Simultaneously, Rousseeuw and Yohai were also working on another alternative for M-estimators, namely *S-estimators* (1984). S-estimators showed to be an improvement on M-estimators' vulnerability to larger levels of contamination and leverage points. A downside, however, is its low efficiency under normal errors.

Using the theory on S-estimators, Yohai (1987) introduced the estimator that, still at the present time, can be considered the default robust linear regression estimator in the literature: Using the best of both M- and S-estimators, Yohai combined the two methods to create *MM-estimators*, benefiting from both high efficiency as well as high breakdown properties. He shows that the MM-estimator does not have a bounded influence curve in the formal definition, but that it is a bounded function for realistic scenarios. Yohai also compared the KW-estimator to the MM-estimator (under normal distributions) and concluded MM-estimators may be better than KW-estimates in the presence of larger levels of contamination, especially in cases with more independent variables. A similar conclusion is drawn from a data example, where KW-estimates are very similar to the (known to be biased) OLS-estimates. An (interesting) overview of the relation between number of covariates p, contamination level ε and the gross error sensitivity γ^* (see Section 3.5.2) for the two estimators is presented in its Table 1.

Lastly, to improve the robustness of maximum likelihood estimators, Cantoni and Ronchetti (2001) developed a robust method based on the quasi-likelihood, developed for generalised linear models.

2.3.3 Robust location and scale estimates

As illustrated in Chapter 2.1 by Hampel et al. (1986), the arithmetic mean estimator also is not robust to outliers. Instead, they show that, for a normal distribution, the sample median is the most robust estimator. Provided that the sample median is a simple estimator and that is shows extremely robust, there are many location estimates introduced using the median and its features. Some exemplary papers are Siegel (1982), introducing the *repeated median* estimator, where nested medians replace the single mean (Siegel, 1982); the *median* of squares regression (Rousseeuw, 1984), as already touched upon in the previous section; and lastly the (generalised) *medians-of-means* (MOM), as is shown by Hsu and Sabato (2014). Furthermore, the median can also be used as a scale estimate, e.g. the median absolute deviation (MAD) estimator (first promoted by Hampel, 1974), though it shows a low Gaussian efficiency (37%). Rousseeuw and Croux (1993) present two alternatives to the MAD, both with higher Gaussian efficiencies at the same high breakdown point.

Another robust estimator is the Minimum Covariance Determinant (MCD) estimator introduced by Rousseeuw (1985), which is a variant of an iteratively reweighted mean with weights determined by the observations' Mahalanobis distance to the current estimate. Though the MCD estimator has a breakdown point of up to 50% as well, it has a tendency to underestimate the variance (excludes too many "good" observations), it only works for symmetric distributions and is can be computationally expensive for larger data sets.

3 Theory

This section further explains important theories and concepts that lie at the foundation of this thesis or provide important distinctions useful for next sections. This section starts with a closer analysis of treatment effect estimation and RCTs (Section 3.1). Next, different type of outliers are defined in Section 3.2 and the term *breakdown point* is defined in Section 3.3. Then, details regarding regression inference, with a focus on the Lin (2013) estimator, is provided in Section 3.4. Lastly, Section 3.5 explains the concepts behind IFs and CVFs further: focusing on the steps of derivation, showing how an estimator's robustness can be evaluated, and providing their general expressions for M-estimators.

3.1 Treatment effect estimation & randomisation

In treatment effect analysis, we want to evaluate the difference in outcome for an individual when treated versus when not treated: $E[Y_i|T_i = 1] - E[Y_i|T_i = 0] = E[Y_{i1} - Y_{i0}]$. An immediate problem arises, namely that the *counterfactual* cannot be observed; an individual is treated ($T_i = 1$) or not ($T_i = 0$) , and thus only one of those terms can be observed. Often, the observable outcome variable Y_i is describes as $Y_i = T_iY_{i1} + (1 - T_i)Y_{i0}$, where T_i is the treatment status indicator and Y_{it} are the respective outcomes when treated ($t = 1$) or untreated ($t = 0$). This phenomenon makes the individual treatment effect unobservable, as was first laid out in the Neyman-Rubin causal model. However, through randomisation of the treatment T_i, the difference in means of both groups is an unbiased estimator for the average treatment effect (ATE).

To mathematically illustrate this statement, the following linear example is provided, echoing Deaton and Cartwright (2018): Let us evaluate a linear causal model of the form:

$$Y_i = \beta_i T_i + \sum_{j=1}^{J} \gamma_j x_{ij},$$

where Y_i is the outcome variable, β_i the individual treatment effect for treatment dummy T_i. When evaluating the difference in means, the following holds:

$$\bar{Y}_1 - \bar{Y}_0 = \bar{\beta}_1 + \sum_{j=1}^{J} \gamma_j (\bar{x}_{1ij} - \bar{x}_{0ij}) = \bar{\beta}_1 - (\bar{S}_1 - \bar{S}_0).$$

The first term on the far-right part of the equation is the often desired *average treatment effect* (ATE) ($\bar{\beta}_1$) in the trial sample and the second term represents the difference in sum of the net average balance of other covariates across the two groups. Now, even with randomisation, it is unlikely that $\bar{S}_1 = \bar{S}_0$ precisely and it is more likely that one is (slightly) larger than the other. However, when the analysis is repeated an infinite amount of times and we additionally assume that there is no post-randomisation correlation of the x-variables with Y, randomisation guarantees that this second term is asymptotically equal to 0. Likewise, the average of the estimated ATEs converges to the true ATE in the trial sample. Interestingly, this holds for unobserved covariates as well, under the assumption of no post-randomisation correlation with covariates occurs.

3.2 Outlier classification

So far, this thesis has addressed outliers as there is only one kind. However, following Rousseeuw and Leroy (1987), outliers can actually be categorised into three categories: *vertical outliers*, *good leverage points* and *bad leverage points*. A visualisation in a simple linear regression context is provided in Figure 1.

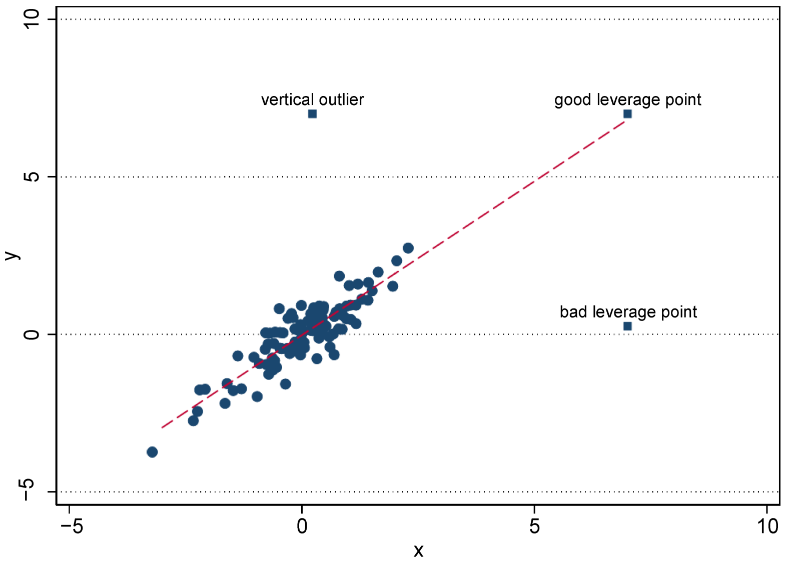


Figure 1: Classification of outliers. Source: Verardi and Croux (2009).

Firstly, vertical outliers are those observations that fall into the “regular” range of the independent (explanatory) variable x, but strongly differ in terms of the dependent (outcome) variable y. In linear regression, vertical outliers can affect the intercept estimation significantly, and can further mildly influence the regression slope coefficients. Secondly, bad leverage points are outlying with respect to the explanatory variables. Their outcome variable can lie either within the normal dependent variable range (as in Figure 1) or deviate even further from the bulk of the data. Often, they have a strong effect on both the intercept and the slope of the estimation. Generally speaking, both vertical outliers and bad leverage points lead to an increase in the variance estimates, as they lie far away from the estimated regression line. Lastly, good leverage points are outlying in both the dimensions of the explanatory and outcome variable, but in a way that they are in line with the regression estimate created using the bulk of observations. They therefore do not affect the regression estimate, but can *decrease* the estimated variance, leading to overoptimistic conclusions about estimate certainty.

3.3 Breakdown point

The idea of an estimator’s breakdown point (concept first introduced by Hampel, 1971) is an easily interpretable concept within robust statistics. An estimator’s breakdown point relates closely to its influence function (Hampel, 1968), as it also aids in understanding the robustness properties of estimators. However, a breakdown point is often investigated for higher percentages, whereas influence functions look at infinitesimal levels of contamination. In the finite sample version of Donoho and Huber (1983), they describe a breakdown point as the "smallest amount of contamination that may cause an estimator to take on arbitrarily large aberrant values". When discussing breakdown points, the contamination always is the worst possible kind (e.g. all outliers lie on one end and all approach infinity (in case for the median estimator)). This also means that contamination levels below the breakdown point give reliable estimates, given an generally well performing estimator. Non-robust estimators have a breakdown point of 0% and the highest attainable level is 50%. An example of an estimator with a breakdown point of α with $0 < \alpha < 50$ is the α-trimmed mean (Hampel, 1974).

Moreover, a difference between breakdown point analysis and influence function analysis, is the level and type of contamination: In the former, one can test larger ε values and the type of contamination is the most adverse contamination possible (for that specific estimator), whereas the latter is performed at $\varepsilon \to 0$ and relates to any contamination other than assumed distribution F (e.g. another distribution or point-mass contamination).

3.4 Coefficient interpretation Lin (2013) estimator

This section provides background knowledge about regression coefficient interpretation, specifically for the Lin (2013) estimator. This section illustrates how the demeaning of the covariates of the interaction term does not introduce additional bias and eases the regression interpretation, through comparing the differences in coefficient interpretation between demeaning the x-variables in the interaction terms versus non-demeaning for the Lin (2013) estimator.

Firstly, it is important to note that the interaction term is added to the regression to accommodate for heterogeneous effect of β on y across the treatment and control group. The estimator/linear regression introduced by Lin (2013) is as follows, including an interaction term where the explanatory variables are demeaned:

$$y = \hat{\mu} + \hat{\tau}T + x\hat{\beta} + (x - \bar{x})T\hat{\gamma} + \hat{\varepsilon}. \tag{1}$$

In linear regression analysis, one is interested in the average marginal effect of a variable, in this application the treatment indicator T, specifically. In the above equation, the marginal effect of T on y is estimated $\hat{\tau} + (x_i - \bar{x})\hat{\gamma}$, making the average marginal effect of T equal to $\hat{\tau}$, because the marginal effect's second term equals 0, on average. This means that the average marginal treatment effect simply equals the coefficient of T. Thus, despite the fact that $\bar{x}$ is not robust to outliers, this term is canceled out, on average. This holds for any value $\bar{x}$ and $\hat{\gamma}$.

To understand the effect of not demeaning x in the interaction term, equation (1) can be rewritten as below by shuffling and regrouping the terms:

$$y = \hat{\mu} + (\hat{\tau} - \bar{x}\hat{\gamma})T + x\hat{\beta} + xT\hat{\gamma} + \hat{\varepsilon}.$$

Now, the formula can be easily compared to the set-up without demeaning:

$$y = \hat{\hat{\mu}} + \hat{\hat{\tau}}T + x\hat{\hat{\beta}} + xT\hat{\hat{\gamma}} + \hat{\hat{\varepsilon}}, \tag{2}$$

We see that the T coefficient has changed ($\hat{\hat{\tau}} = \hat{\tau} - \bar{x}\hat{\gamma}$) and all other terms remain unchanged. When evaluating the not-demeaned alternative (2), we find that the marginal effect of T is $\hat{\hat{\tau}} + x_i\hat{\hat{\gamma}}$, making the average marginal effect $\hat{\hat{\tau}} + \bar{x}\hat{\hat{\gamma}}$. This means that in the not-demeaned case, the bias of $\bar{x}$ is introduced into the analysis, instead. Additionally, the analysis is easier to conduct in the demeaned case, as the inference can be directly performed from the linear regression output by any statistical software. Obviously, outliers can still effect all estimator estimates (e.g. OLS estimates are biased in the presence of bad leverage points), but they could be robustly estimated using a robust method like an MM-estimator.

3.5 The influence- & change-of-variance function

First introduced by Hampel (1968, 1974), the influence function is defined as the standardised bias on the estimator because of infinitesimal point-mass contamination ε at point $z \in \mathbb{R}$. Mathematical representation of the influence function of estimator T at underlying distribution F is defined as

$$IF(z; T, F) = \lim_{\varepsilon \to 0} \big\{T(F_\varepsilon) - T(F)\big\}/\varepsilon,$$

where $F_\varepsilon = (1 - \varepsilon)F + \varepsilon\Delta_z$ and Δ_z denotes the point mass contamination at any point $z \in \mathbb{R}$ (Tukey-Huber contamination model). Similarly, influence functions can also be described as the marginal effect of contamination on the estimator's derivative:

$$IF(z; T, F) = \frac{\partial}{\partial\varepsilon} T(F_\varepsilon)\Big|_{\varepsilon=0}. \tag{3}$$

For the next parts, notation is as follows: Let F_N indicate the empirical distribution of observations z and let F be the true distribution of z. F_ε indicates the contaminated distribution. Furthermore, let T be the parameter functional such that $T(F) = \tau$ and $T(F_N) = \hat{\tau}$ are the treatment effect and its estimator, respectively.

3.5.1 Derivational steps

The derivation of an estimator' IF starts with its *score functions.* A score function Ψ is the derivative of an estimator's *loss function* ρ. Giving a continuously piece-wise differentiable loss function ρ, total loss can be minimised by differentiating ρ with respect to estimator T and equating it to 0. Then, score functions are such that

$$\frac{\partial}{\partial T}\rho(z;T,F) = \Psi(z;T,F) = 0.$$

From this, we can consider the following equation:

$$E_F\Big[\Psi\big\{z;T(F),F\big\}\Big] = 0.$$

for the super population. After introducing distributional contamination $F \to F_\varepsilon$, $T(F_\varepsilon)$ is defined by solving:

$$\int \Psi\big\{z;T(F_\varepsilon),F_\varepsilon\big\}dF_\varepsilon = 0.$$

As the influence function is defined by the partial derivative of $T(F_\varepsilon)$ with respect to ε at the limit $\varepsilon \to 0$, the influence function is found by solving

$$\frac{\partial}{\partial\varepsilon}\int \Psi\big\{z;T(F_\varepsilon),F_\varepsilon\big\}dF_\varepsilon\Big|_{\varepsilon=0} = 0$$

for $\frac{\partial}{\partial\varepsilon}T(F_\varepsilon)|_{\varepsilon=0} = IF(z;T,F)$ (equation (3)). Moreover, when evaluating robustness, it is interesting to not only evaluate the effect on the estimator's asymptotic value but also to evaluate its asymptotic variance. This introduces the next fundamental element of robustness, namely the change-of-variance function.

It is worth noting that the IF and CVF show many similarities, for example $\int IF(z;T,F)dF = \int CVF(z;T,F)dF = 0$. These characteristics are important for the derivations of the curves, as evident in the next section. The CVF is derived as follows:

$$CVF(z;T,F) = \frac{\partial}{\partial\varepsilon}V(T,F_\varepsilon)\Big|_{\varepsilon=0} = \frac{\partial}{\partial\varepsilon}V\big\{T,(1-\varepsilon)F+\varepsilon\Delta_z\big\}\Big|_{\varepsilon=0}.$$

3.5.2 Robustness evaluation

After deriving the influence function, estimator robustness can be evaluated. Following Hampel et al. (1986), there are three measures for evaluating robustness: *the gross-error sensitivity*, *the local-shift sensitivity*, and the *rejection point.* The first type of robustness is the focal point of this thesis and is also considered the most important one within the literature. However, the local-shift sensitivity measure is worth mentioning as well, as it is characteristic for the median estimator, as it relates to the "wiggling" of an estimator around the center of symmetry. Hampel et al. (1986) provide a simple comparison between the sample mean and median for all three measures.

The gross-error sensitivity is the supremum over z of the influence function's absolute value:

$$\gamma^*(T,F) = \sup_z |IF(z;T,F)|.$$

It is said that an estimator is bias-robust (B-robust) if the gross-error sensitivity score γ^* is bounded for infinitesimal levels of contamination (Hampel et al., 1986). Simple exemplary IF-applications for the arithmetic mean and robust alternatives are provided in Hampel (1974). Note that for M-estimators, the gross-error sensitivity γ^* is only bounded when score function Ψ is bounded as well (Krasker, 1980), because the IF is proportional to the score function. Moreover, also note that a breakdown point of 0% translates to $\gamma^* \to \infty$.

For the variance, there is a similar measure κ^* for evaluating variance-robustness (V-robust):

$$\kappa^* = \sup_z \big\{CVF(z;T,F)/V(T,F)\big\},$$

This V-robustness is a more stringent robustness concept than B-robustness, and V-robustness implies B-robustness (Rousseeuw, 1981b). An estimator is V-robust if and only if the *change-of-variance sensitivity* κ^* is bounded.

3.5.3 General M-estimator expressions

Though a general understanding of the derivational steps is important, this thesis derives the treatment estimators' IFs and CVFs using Zhelonkin (2013)'s work on M-estimators. For one-stage M-estimators, the IF is proportional to the score function (see Appendix A.1 for the derivation), which means that the score function must be bounded for the IF to be bounded. In turn, this means that an estimator's value is robust to outliers if and only if its score function is bounded. For a one-stage M-estimator, the influence function is as follows:

$$IF(z;S,F) = M^{-1}\Psi_1\big\{z^{(1)}, S(F)\big\},$$

where $M = -\int \frac{\partial}{\partial\theta}\Psi_1(z,\theta)dF$ and Ψ_1 represents the score function. Moreover, for M-estimators, the asymptotic variance can then be expressed as:

$$V(S,F) = \int IF(z;S,F)IF(z;S,F)^T dF, \tag{4}$$

given some regularity conditions and symmetric distribution F (Hampel et al., 1986). For one-stage M-estimators specifically, this resembles the following structure:

$$\begin{aligned} V(S,F) =& \int IF(z;S,F)IF(z;S,F)^T dF \\ =& M^{-1}\int \Psi_1\big\{z^{(1)}, S(F)\big\}\Psi_1\big\{z^{(1)}, S(F)\big\}^T dF M^{-1}. \end{aligned} \tag{5}$$

The derivation for the one-stage CVF is also included in Zhelonkin (2013)'s work, but is omitted in this thesis.

For two-stage M-estimators, the IF, variance and CVF can also be derived. For the IF, it extends the one-stage IF estimator to:

$$
\begin{aligned}
IF(z;T,F) =& M^{-1}\Bigg(\Psi_2\Big[z^{(2)}, h\big\{z^{(1)}, S(F)\big\}, T(F)\Big] \\
&+\int \frac{\partial}{\partial\theta}\Psi_2\big\{\tilde{z}^{(2)}, \theta, T(F)\big\}\frac{\partial}{\partial\eta}h(\tilde{z}^{(1)}, \eta)dF \cdot IF(z;S,F)\Bigg),
\end{aligned}
\tag{6}
$$

where $M = -\int \frac{\partial}{\partial\xi}\Psi_2\Big[\tilde{z}^{(2)}, h\big\{\tilde{z}^{(1)}, S(F)\big\}, \xi\Big]dF$.

The two-stage asymptotic variance can be described as, again following (4), now using $IF(z;T,F)$ instead:

$$
V(T,F) = M^{-1}\int a(z)a(z)^T + a(z)b(z)^T + b(z)a(z)^T + b(z)b(z)^T dF M^{-1}, \tag{7}
$$

where $a(z) = \Psi_2\Big[z^{(2)}, h\big\{z^{(1)}, S(F)\big\}, T(F)\Big]$ and $b(z) = \int \frac{\partial}{\partial\theta}\Psi_2\Big[z^{(2)}, h\big\{z^{(1)}, S(F)\big\}, T(F)\Big] \cdot \frac{\partial}{\partial\eta}h(z^{(1)}, \eta) \cdot IF(z;S,F)$. Moreover, Zhelonkin (2013) uses the above expressions to also derive a general expression for the two-stage M-estimator's CVF as well:

$$
\begin{aligned}
CVF(z;S,T,F) = & V(T,F) - M^{-1}\Big(\int D^{(2S)}dF\Big)V(T,F) \\
& - M^{-1}\Big(\frac{\partial}{\partial\theta}\Psi_2\Big[z^{(2)}, h\big\{z^{(1)}, S(F)\big\}, \theta\Big]\Big)V(T,F) \\
& + M^{-1}\Big(\int \big\{Aa(z)^T + Ba(z)^T + Ab(z) + Bb(z)^T\big\}dF\Big)M^{-1} \\
& + M^{-1}\Big(\int \big\{a(z)A^T + b(z)A^T + a(z)B^T + b(z)B^T\big\}dF\Big)M^{-1} \\
& + M^{-1}\Big(a(z)a(z)^T + a(z)b(z)^T + b(z)a(z)^T + b(z)b(z)^T\Big)M^{-1} \\
& - V(T,F)\Big(\int D^{(2S)}dF + \frac{\partial}{\partial\theta}\Psi_2\Big[z^{(2)}, h\big\{z^{(1)}, S(F)\big\}, \theta\Big]\Big)M^{-1},
\end{aligned}
\tag{8}
$$

where $D^{(2S)}$ is a matrix with elements

$$
\begin{aligned}
D^{(2S)}_{ij} =& \Big(\frac{\partial}{\partial h}\frac{\partial\Psi_{2i}(z^{(2)}, h, \theta)}{\partial\theta_j}\Big)^T \frac{\partial h(z^{(1)}, s)}{\partial s} IF(z;S,F) \\
&+ \Big(\frac{\partial}{\partial\theta}\frac{\partial\Psi_{2i}(z^{(2)}, h, \theta)}{\partial\theta_j}\Big)^T IF(z;T,F),
\end{aligned}
$$

matrix A is given by

$$
\begin{aligned}
A =& \frac{\partial}{\partial h}\Psi_2\big\{z^{(2)}, h, T(F)\big\}\frac{\partial h(z^{(1)}, s)}{\partial s} IF(z;S,F) \\
&+ \frac{\partial}{\partial\theta}\Psi_2\Big[z^{(2)}, h\big\{z^{(1)}, S(F)\big\}, \theta\Big] IF(z;T,F),
\end{aligned}
$$

and matrix B has the form

$$
\begin{aligned}
B =& \int R_1\frac{\partial}{\partial s}h(z^{(1)}, s)dF IF(z;S,F) + \int \frac{\partial}{\partial h}\Psi_2\big\{z^{(2)}, h, T(F)\big\}R_2 dF IF(z;S,F) \\
&- \int \frac{\partial}{\partial h}\Psi_2\big\{z^{(2)}, h, T(F)\big\}\frac{\partial}{\partial s}h(z^{(1)}, s)dF M_1^{-1}\int D^{(1)}dF IF(z;S,F) \\
&+ \frac{\partial}{\partial h}\Psi_2\big\{z^{(2)}, h, T(F)\big\}\frac{\partial}{\partial s}h(z^{(1)}, s) IF(z;S,F),
\end{aligned}
$$

where matrix $D^{(1)}$ has elements

$$D^{(1)}_{ij} = \Big(\frac{\partial}{\partial\theta}\frac{\partial\Psi_{1i}(z^{(1)},\theta)}{\partial\theta_j}\Big)^T IF(z;S,F);$$

and matrix $R^{(1)}$ has elements

$$\begin{aligned} R^{(1)}_{ij} =& \Big(\frac{\partial}{\partial h}\frac{\partial\Psi_{2i}(z^{(2)},h,T(F))}{\partial h_j}\Big)^T + \frac{\partial h(z^{(1)},s)}{\partial s} IF(z;S,F) \\ &+ \frac{\partial}{\partial\theta}\frac{\partial\Psi_{2i}(z^{(2)},h,\theta)}{\partial h_j} IF(z;T,F), \end{aligned}$$

and $R^{(2)}$ is a matrix with elements $R^{(2)}_{ij} = \Big(\frac{\partial}{\partial s}\frac{\partial h_i(z^{(1)},s)}{\partial s_j}\Big)^T IF(z;S,F)$, and M_1 indicates the first-stage's M matrix. Furthermore, this thesis' notation of matrix B is slightly different from matrix B in Zhelonkin (2013), because two terms included cancel each other out perfectly. For those interested, detailed derivations can be found in Zhelonkin (2013)'s Appendix section, specifically sections A.3 and A.4.

4 Robustness Evaluation

In this section, the IF and CVF are derived and discussed for each of the three treatment effect methods. The derivations closely follow the work of Zhelonkin (2013), specifically its Sections 3.1-3.4 and appendices A.3 and A.4. This section is structured as follows: Firstly, an overview of the different treatment effect estimators is provided (Section 4.1). Secondly, the IF and CVF for each of the estimators are provided in subsections 4.2-4.4. Again, robustness is evaluated for infinitesimal levels of contamination using the Tukey-Huber contamination model $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_z$. Thirdly, an analysis of their robustness to the different types of outliers is performed and more robust alternatives are suggested, which form the base for this thesis' methodology (Section 6).

4.1 Overview estimators

This thesis evaluates three different treatment effect estimators. The first two estimators that are discussed are one-stage estimators, whereas the third is a two-stage estimator. These three estimators are the most present estimators in the literature: the first estimator is the classical, unbiased difference-in-means estimator; the second estimator adjusts the first one by including covariates and is known to have a lower variance in clean samples; and the third estimator is the two-step equivalent of the second estimator. All three estimators are M-estimators, allowing the use of general M-estimator derivations as presented in Section 3.5.3. An overview of the estimators and corresponding score functions (using squared loss functions) is provided in Table 1 and detailed derivations can be found in Appendix A.

Table 1: Overview of the different treatment effect estimators.

Difference-in-means (DiM)	
Estimator ($\hat{\tau}$)	$\hat{\tau} = \frac{1}{N_1}\sum_{i \in T} y_i - \frac{1}{N_0}\sum_{i \in C} y_i$
$\Psi_1\big\{z^{(1)}, S(F)\big\}$	$\begin{bmatrix} y_1 - \mu_1 \\ y_0 - \mu_0 \end{bmatrix}$, with $z^{(1)} = \{y_1, y_0\}$ and $S(F) = \begin{bmatrix} \mu_1 \\ \mu_0 \end{bmatrix}$
$\Psi_2\Big[z^{(2)}, h\big\{z^{(1)}, S(F)\big\}, T(F)\Big]$	$\mu_1 - \mu_0 - \tau = S(F)^T \begin{bmatrix} 1 \\ -1 \end{bmatrix} - \tau$, with $z^{(2)}$ empty; $h\big\{z^{(1)}, S(F)\big\} = \mu_1 - \mu_0$ and $T(F) = \tau$
Regression adjusted (RA)	
Estimator ($\hat{\tau}$)	$y = \hat{\mu} + \hat{\tau} T + x^T \hat{\beta} + T(x - \bar{x})^T \hat{\gamma} + \hat{\varepsilon}$
$\Psi_1\big\{z^{(1)}, S(F)\big\}$	$\big(y - \mu - \tilde{\tau} T - x^T \beta - T(x - \mu_x)^T \gamma\big) \begin{bmatrix} 1 \\ T \\ x \\ T(x - \mu_x) \end{bmatrix}$, with $z^{(1)} = \{y, x, T\}$ and $S(F) = \begin{bmatrix} \mu & \tau & \beta^T & \gamma^T \end{bmatrix}$
$\Psi_2\Big[z^{(2)}, h\big\{z^{(1)}, S(F)\big\}, T(F)\Big]$	$\tilde{\tau} - \tau = S(F)^T \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} - \tau$, with $z^{(2)}$ empty; $h\big\{z^{(1)}, S(F)\big\} = \tilde{\tau}$ and $T(F) = \tau$
Difference-in-intercepts (DiI)	
Estimator ($\hat{\tau}$)	Step 1: $y_i = \hat{\mu}_T + (x_i - \bar{x})^T \hat{\beta}_T + \hat{\varepsilon}_i$ for $i \in T = 0, 1$ Step 2: $\hat{\tau} = \hat{\mu}_1 - \hat{\mu}_0$
$\Psi_1\big\{z^{(1)}, S(F)\big\}$	$\begin{bmatrix} \big(y_1 - \mu_1 - (x_1 - \mu_{x1})^T \beta_1\big) \begin{bmatrix} 1 \\ x_1 - \mu_{x1} \end{bmatrix} \\ \big(y_0 - \mu_0 - (x_0 - \mu_{x0})^T \beta_0\big) \begin{bmatrix} 1 \\ x_0 - \mu_{x0} \end{bmatrix} \end{bmatrix}$, with $z^{(1)} = \{y_1, x_1, y_0, x_0\}$ and $S(F) = \begin{bmatrix} \mu_1 & \beta_1^T & \mu_0 & \beta_0^T \end{bmatrix}$
$\Psi_2\Big[z^{(2)}, h\big\{z^{(1)}, S(F)\big\}, T(F)\Big]$	$\mu_1 - \mu_0 - \tau = S(F)^T \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix} - \tau$, with $z^{(2)}$ empty; $h\big\{z^{(1)}, S(F)\big\} = \mu_1 - \mu_0$ and $T(F) = \tau$

4.2 Difference-in-means estimator

The difference-in-means (DiM) estimator is an unbiased estimator for the average treatment effect. It is powerful in its simplicity and easy to compute. As can also be found in Table 1, the estimator is as follows:

$$\hat{\tau} = \frac{1}{N_1}\sum_{i\in T}^{N_1} y_i - \frac{1}{N_0}\sum_{i\in C}^{N_0} y_i$$

Though the estimator is a one-step estimator, its IF is constructed using two score functions: The first-stage score function is there to compute the (arithmetic) means (i.e. two mean IF functions stacked on top of each other) and the second one is for computing the final estimator (i.e. subtraction operator). Appendix A.2 shows the detailed derivations for this estimator. Both score functions are derived using a squared loss function and can be found below:

$$\Psi_1\{z^{(1)}, S(F)\} = \begin{bmatrix} y_1 - \mu_1 \\ y_0 - \mu_0 \end{bmatrix}$$

$$\Psi_2\Big[z^{(2)}, h\{z^{(1)}, S(F)\}, T(F)\Big] = \mu_1 - \mu_0 - \tau = S(F)^T \begin{bmatrix} 1 \\ -1 \end{bmatrix} - \tau,$$

where $z^{(1)} = \{y_1, y_0\}$, $S(F) = \begin{bmatrix} \mu_1 \\ \mu_0 \end{bmatrix}$, $z^{(2)}$ is empty, $h\{z^{(1)}, S(F)\} = \mu_1 - \mu_0 = S(F)^T \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ and $T(F) = \tau$. Also note that Ψ_1 is a 2×1-vector and that Ψ_2 is a scalar.

Using the two-stage M-estimator IF formula (6), it is found that

$$IF(z; T, F) = y_1 - y_0 - \tau, \tag{9}$$

where we use that $\frac{\partial \Psi_2}{\partial h} = 1$, $\frac{\partial h(z^{(1)},\theta)}{\theta} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ and $M = 1$. It is evident that the estimator is not (value-)robust to outliers in either the treatment or control group, as they directly impact the influence function.

The CVF can be computed using the same elements and formula (8). The difference-in-means CVF is

$$CVF(z; S, T, F) = 3V(T, F) + 2\int \{a(z)A + b(z)^2\}dF + \{a(z) + b(z)\}^2, \tag{10}$$

where $a(z) = -A = \Psi_2$ and $b(z) = \begin{bmatrix} 1 & -1 \end{bmatrix} \Psi_1 = \begin{bmatrix} 1 & -1 \end{bmatrix} IF(z; S, F)$. Please note that $IF(z; S, F) = \Psi_1\{z^{(1)}, S(F)\} = \begin{bmatrix} y_1 - \mu_1 \\ y_0 - \mu_0 \end{bmatrix}$, as $M_1 = 1$. The detailed CVF derivation can be found in Appendix A.3. It is evident that the boundedness of the CVF is determined by the boundedness of $IF(z; T, F)$ through $V(T, F)$ (see asymptotical variance expression (4)) and $IF(z; S, F)/\Psi_1$ through $b(z)$. Note that Ψ_2 is a constant, therefore making $a(z)/A$ bounded and making $IF(z; T, F)$ unbounded only through Ψ_1.

Evaluating robustness properties, it is evident that this estimator is neither V-robust nor B-robust as IF and CVF are both unbounded. This is due to its *fragile* score function, explaining why the arithmetic means is not robust to infinitesimal levels of data perturbation. Specifically, outliers in both the treatment and control group can draw the arithmetic mean $\bar{y} \to (-)\infty$, inflating both the IF and CVF of the estimator to ∞ as well. The estimator would be robust to infinitesimal levels of contamination with a bounded Ψ_1. This thesis addresses two alternative estimators, replacing the arithmetic means: the median and the α-trimmed mean.

4.2.1 Robust alternatives

The first robust estimator to consider is the difference-in-medians (DiMed) estimator:

$$\hat{\tau} = y_1^{(N_1+1)/2} - y_0^{(N_0+1)/2},$$

where, for both the treatment and control group, y are ordered observations, such that $y^{(1)}$ and $y^{(N)}$ are the minimum and maximum of y, respectively. The corresponding score function is

$$\Psi_{1,DiMed}\{z^{(1)}, S(F)\} = \begin{bmatrix} -\frac{y_1-\eta_1}{|y_1-\eta_1|} \\ -\frac{y_0-\eta_0}{|y_0-\eta_0|} \end{bmatrix},$$

where η represents the median. The score function is derived from the absolute loss function $\rho = |y - \eta|$, by taking the derivative with respect to true median η. Just like for the means, the two score functions for the treatment and control group are stacked together into one Ψ_1. It must be noted that the elements in Ψ_1 is not defined at their medians, because then $y = \eta \Leftrightarrow y - \eta = 0$, setting the denominatior to 0 (this explains its local-shift sensitivity). It can easily be shown that the elements in $\Psi_{1,DiMed}$ take on either 1 or -1 (-1 for $y_t > \eta_t$ and $+1$ for $y_t < \eta_t$), and thus is a bounded score function. This also leads to bounded IF_{DiMed}

$$IF_{DiMed}(z; S, F) = \begin{bmatrix} \frac{sign(y_1)}{2f(M)} \\ \frac{sign(y_0)}{2f(M)} \end{bmatrix} \text{(Hampel et al., 1986)},$$

where $M = F^{-1}(0.5)$ (for a normal distribution $f(M) = \phi(0)$). For symmetric distributions and under no contamination, the median is an unbiased estimator for the population mean. Under contamination, however, it must be noted that the median does not remove or down-weight the outliers. Instead, the median estimator moves a little in the direction of the outlier (i.e. the contaminated estimate is now be the initial estimate's neighbouring observation). This also explains its IF and CVF bounded properties: introducing one outlier in the sample will never make the estimator approach infinity, but rather take on another finite value. According to an example by Hampel et al. (1986) in Section 2.5c, the median is the most (V-)robust estimator for a (standard) normal distribution. In terms of CVF, the difference-in-median CVF equals that of the difference-in-means CVF structure in (10), now with robust Ψ_1 and $IF_{median}(z; S, F)$ bounded alternatives.

Another alternative for the arithmetic mean estimator is the α-trimmed mean (DiMT). Here, the score function does not change, but $z^{(1)}$ is trimmed. After sorting, the outer most α percentage of all observations (within each group) are removed from the sample. If the estimator is successful in perfectly removing all outliers, the estimator remains unbiased and the variance decreases. Moreover, removing outer observations from (clean) thicker-tailed distributions (e.g. Cauchy) leads to a decrease in the variance estimate, whereas the trimmed mean is a less efficient estimator (compared to the basic arithmetic mean) in the thinner-tailed normal distribution. Furthermore, decreasing the trimming percentage converges the estimator towards the difference-in-means estimator and increasing it converges the estimator towards the difference-in-medians estimator, for both value and variance.

It must be noted that this alternative does not have a bounded IF or CVF function. This is because outliers are only removed if they fall in the α-percentage tails of the sample. Rather, the IF and CVF represent the effects of infinitesimal levels of contamination anywhere in the sample (i.e. in this thesis, at any point-mass Δ_z). Still, this can be an interesting estimator to consider for practical applications, and is therefore included in the analysis.

4.3 One-stage regression-adjusted estimator

A way to extend the difference-in-means estimator is through including covariates with predictive power to the analysis. Not only the covariates, but also an interaction term between the covariates and the binary treatment variable is added to the regression model, such that it can deal with possible heterogeneity. This estimator is called the Regression-Adjusted OLS-estimator (RA OLS). Copying Table 1, the estimator $\hat{\tau}$ is obtained from the OLS-regression

$$y = \hat{\mu} + \hat{\tau}T + x^T\hat{\beta} + T(x - \hat{\mu}_x)^T\hat{\gamma} + \hat{\varepsilon}.$$

Similar to the difference-in-means estimators, despite the estimator itself being a one-stage estimator, there are two score functions. The first one is to estimate all coefficients in the regression:

$$\Psi_1\big\{z^{(1)}, S(F)\big\} = \big(y - \mu - \tilde{\tau}T - x^T\beta - T(x - \mu_x)^T\gamma\big)\begin{bmatrix} 1 \\ T \\ x \\ T(x - \mu_x) \end{bmatrix}$$

with $z^{(1)} = \{y, x, T\}$ and $S(F)^T = \begin{bmatrix}\mu & \tilde{\tau} & \beta^T & \gamma^T\end{bmatrix}$ and the second score function is to attain the coefficient matching the treatment indicator variable:

$$\Psi_2\Big[z^{(2)}, h\big\{z^{(1)}, S(F)\big\}, T(F)\Big] = \tilde{\tau} - \tau = S(F)^T\begin{bmatrix}0\\1\\0\\0\end{bmatrix} - \tau$$

with $z^{(2)}$ empty, $h\big\{z^{(1)}, S(F)\big\} = \tilde{\tau} = S(F)^T\begin{bmatrix}0\\1\\0\\0\end{bmatrix}$ and $T(F) = \tau$. Please note that the treatment indicator coefficient $\tilde{\tau}$ in Ψ_1 has been given a tilde, to be able to differentiate τ between the two score functions (the true treatment effect estimator in Ψ_2 is also named τ, following notation in the other estimators). Also note that Ψ_1 is a $2(p+1) \times 1$-vector with p equal to number of covariates and that Ψ_2 is a scalar.

This estimator is a two-stage M-estimator, where we get its IF using (6):

$$IF(z; T, F) = \Psi_2\Big[z^{(2)}, h\big\{z^{(1)}, S(F)\big\}, T(F)\Big] + \begin{bmatrix}0\\1\\0\\0\end{bmatrix}^T M_1^{-1}\Psi_1\big\{z^{(1)}, S(F)\big\},$$

where we use that $\frac{\partial \Psi_2}{\partial h} = 1$, $\frac{\partial h(z^{(1)},\theta)}{\theta} = \begin{bmatrix} 0 & 1 & 0^T & 0^T \end{bmatrix}$, $M = 1$ and the first-stage influence function $IF(z; S, F) = M_1^{-1}\Psi_1\{z^{(1)}, S(F)\}$. These elements can be used to derive the asymptotic variance, using (4):

$$
\begin{aligned}
V(T,F) &= \int IF(z;T,F)IF(z;T,F)^T dF \\
&= \int \Psi_2^2 dF + 2 \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}^T M_1^{-1} \int \Psi_1\{z^{(1)}, S(F)\} dF \\
&+ \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}^T \int M_1^{-1}\Psi_1\{z^{(1)}, S(F)\}\Psi_1\{z^{(1)}, S(F)\}^T M_1^{-1} dF \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} 0 & 1 & 0^T & 0^T \end{bmatrix} Var(S,F) \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} = Var(\tilde{\tau}),
\end{aligned}
\tag{11}
$$

where we used that $\int \Psi_2 dF = \int \Psi_1 dF = 0$ and the expression for the first-stage variance $V(S, F)$, as seen in equation (7). Using the two-stage M-estimator CVF function (equation (8)), this gives a CVF formula of the same structure as the difference-in-means CVF (equation (10)), now with $a(z) = \Psi_2$ and $b(z) = \begin{bmatrix} 0 & 1 & 0^T & 0^T \end{bmatrix} IF(z; S, F)$. Notes for the CVF derivation can be found in Appendix A.4. Again, the unboundedness of the CVF depends on the unboundedness of $IF(z; S, F)/\Psi_1$.

When analysing the estimator's (B-)robustness, we can see that the total effect of $IF(z; S, F)$ on $IF(z; T, F)$ is restricted to only its second row (out of $2(p+1)$ rows). This second line equals the following: $\big(y - \mu - \tilde{\tau}T - x^T\beta - T(x - \mu_x)^T\gamma\big) \cdot T$ (for convenience, this is denoted as $\Psi_{1,2}$). In this thesis, we assume no misspecification of treatment status, thus the second term T cannot be an area of concern. The former term, however, resembles the error term between the true outcome variable and the expected value. With vertical outliers, the x-value does not deviate much from μ_x, but outcome variable y deviates much, therefore causing $\Psi_{1,2} \to \infty$. Depending on whether the outlier lies in the treatment or control group, either $\hat{\tau}$ changes (treatment group) or both $\hat{\tau}$ and $\hat{\mu}$ (control group) are affected. This is because the line of estimation shifts vertically towards the outlier, relatively uneffecting its slope coefficients. For bad leverage points, x and y both deviate from the regression line, making the residual large. In the worst case, $\Psi_{1,2} \to \infty$. In turn, this outlier can draw the estimation line towards itself to minimise the otherwise (even) larger residual. Bad leverage points in the control group can alter all estimators $\hat{\mu}$, $\hat{\tau}$, $\hat{\beta}$, and $\hat{\gamma}$ and in the treatment group only $\hat{\tau}$ and $\hat{\gamma}$ are expected to change significantly. For good leverage points, we see that the residual is low, and can therefore draw $\Psi_{1,2}$ down to 0, making $IF(z; T, F)$ finite and nearing 0 as $\Psi_2 \to 0$.

Similar effects are seen for the variance and CVF: As evident from (11), the variance is restricted to only the variance from the treatment indicator coefficient; other estimators' variances are not included, as desired. For vertical outliers, $IF(z; T, F)$ and $IF(z; S, F) \to \infty$, therefore also inflating $V(T, F)$ and $b(z)$. Clearly, the CVF is not bounded for vertical outliers. The same holds for bad leverage points. For good leverage points, the overall variance decreases slightly (more for each additional good leverage point). A similar effect can be seen in the CVF, which is larger than 0, but bounded. Again, good leverage points have a bounded effect in the sense that the change-of-variance sensitivity κ^* is bounded, but they can cause the variance (and CVF) to "falsely" move downwards.

4.3.1 Robust alternatives

This thesis discusses two robust alternatives: The first one, is Yohai (1987)'s MM-estimator with Tukey biweight loss functions (Tukey, 1977), provided that it is the default estimator in the literature and has a high breakdown point. The second is the KW-estimator (Krasker 1980; Krasker and Welsch 1982), since it is the most V-robust estimator in the literature and the focus of this thesis lies at infinitesimal levels of contamination.

For the MM-estimator, tuning parameters $c_0 = 1.548$ and $c_1 = 4.685$ are applied to the Tukey biweight loss functions. A formal derivation of the IF and CVF using an MM-estimator lies beyond the scope of this thesis, but Yohai (1987) shows that a consistent estimator with a 50% breakdown point can be achieved with high efficiency, when errors are normally distributed. Following the regression structure of the OLS alternative, this estimator is called the Regression-adjusted MM-estimator (RA MM). The Tukey loss function is as follows:

$$\rho(r) = \begin{cases} \frac{c^2}{6}\Big(1 - \big[1 - (\frac{r}{c})^2\big]^3\Big), & \text{if } |r| \leq c \\ \frac{c^2}{6}, & \text{otherwise} \end{cases} \tag{12}$$

with score function

$$\Psi(r) = \begin{cases} r\big[1 - (\frac{r}{c})^2\big]^2, & \text{if } |r| \leq c \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

where $r = \frac{u_i}{s}$ is the scaled residual and c is a positive tuning parameter. $c = 1.548$ in the first step to tune for high breakdown and $c = 4.658$ in the second step for high efficiency.

The KW-estimator does have a bounded influence function of the form:

$$IF(z; T, F) = \Psi\{(y - x\beta)A^{-1}x^T\},$$

where A is a $p \times p$ matrix satifying

$$A = E\Big[2\phi(\frac{a}{\sigma|A^{-1}x^T|}) - 1\Big]x^T x.$$

The estimator is a Modified Least Squares (MLS) estimator and is the optimal V-robust estimator for linear regression (Ronchetti and Rousseeuw, 1985). However, it is also shown to have a low breakdown point in multivariate data and it is interesting to compare the results of the MM- and KW-estimator in this a simulation with little contamination. This estimator is called the regression-adjusted KW-estimator (RA KW).

4.4 Two-stage difference-in-intercepts estimator

As mentioned earlier, this estimator is a two-stage estimator, where two individual regressions are run, after which the estimator is the difference between the two intercept estimates. Following Table 1, the covariate coefficients and intercepts for both the treatment and control group can be estimated through (step 1)

$$y_t = \hat{\mu}_t + (x_t - \bar{x}_t)^T \hat{\beta}_t + \hat{\varepsilon}_t \text{ for } t = 0, 1.$$

Then (step 2), the final the difference-in-intercepts (DiI OLS) estimator is

$$\hat{\tau} = \hat{\mu_1} - \hat{\mu_0}.$$

The corresponding two score functions are Ψ_1

$$\Psi_1\{z^{(1)}, S(F)\} = \begin{bmatrix} (y_1 - \mu_1 - (x_1 - \mu_{x1})^T \beta_1) \begin{bmatrix} 1 \\ x_1 - \mu_{x1} \end{bmatrix} \\ (y_0 - \mu_0 - (x_0 - \mu_{x0})^T \beta_0) \begin{bmatrix} 1 \\ x_0 - \mu_{x0} \end{bmatrix} \end{bmatrix},$$

where $z^{(1)} = \{y_1, y_0, x_1, x_0\}$ and $S(F)^T = \begin{bmatrix} \mu_1 & \beta_1^T & \mu_0 & \beta_0^T \end{bmatrix}$ and Ψ_2 is

$$\Psi_2\Big[z^{(2)}, h\big\{z^{(1)}, S(F)\big\}, T(F)\Big] = \mu_1 - \mu_0 - \tau = S(F)^T \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix} - \tau,$$

where $z^{(2)}$ is empty, $h\big\{z^{(1)}, S(F)\big\} = \mu_1 - \mu_0 = S(F)^T \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix}$ and $T(F) = \tau$ Also note that Ψ_1 is a $2(p+1) \times 1$-vector with p equal to number of covariates and that Ψ_2 is a scalar.

Again, using the two-stage M-estimator IF equation (6), the difference-in-intercept IF can be constructed:

$$IF(z; T, F) = \Psi_2\Big[z^{(2)}, h\big\{z^{(1)}, S(F)\big\}, T(F)\Big] + \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix}^T M_1^{-1} \Psi_1\big\{z^{(1)}, S(F)\big\},$$

where we use that $\frac{\partial \Psi_2}{\partial h} = 1$, $\frac{\partial h(z^{(1)}, \theta)}{\theta} = \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix}^T$, $M = 1$ and the first-stage influence function $IF(z; S, F) = M_1^{-1} \Psi_1\{z^{(1)}, S(F)\}$. These elements can be used to derive the asymptotic variance, using (4) and following similar steps as for in (11):

$$V(T, F) = \begin{bmatrix} 1 & 0^T & -1 & 0^T \end{bmatrix} Var(S, F) \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix} = Var(\mu_1) + Var(\mu_0),$$

where we used that $\int \Psi_2 dF = \int \Psi_1 dF = 0$ and the expression for the first-stage variance $V(S, F)$, as seen in equation (7). Using the two-stage M-estimator CVF function (equation (8)), this gives a CVF formula of the same structure as the difference-in-means CVF (equation (10)), now with $a(z) = \Psi_2$ and $b(z) = \begin{bmatrix} 1 & 0^T & -1 & 0^T \end{bmatrix} IF(z; S, F)$. Notes for the CVF derivation can be found in Appendix A.5. Again, the unboundedness of the CVF depends on the unboundedness of $IF(z; S, F)/\Psi_1$.

F

4.4.1 Robust alternative

Similar to the one-stage regression adjusted estimator, robust alternatives are the MM-estimator (DiI MM) and the KW-estimator (DiI KW). For the MM-estimator, the same loss functions and tuning parameters as in the regression-adjusted method are used.

5 Data & experimental set-up

This section describes the data for a simulation study (Section 5.1) and a real-data set application (Section 5.2). The former includes a brief explanation on the experimental set-up regarding number of observations and different type of outliers included, as well.

5.1 Simulation study

A simulation study is conducted to analyse the effects of the three types of outliers on the three value- and variance-estimates. Each measurement is repeated 500 times and a sample sizes of $N = 2000$ is used. The DGP is the following model:

$$\begin{aligned} y_0 &= \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon, \\ y_1 &= y_0 + \tau, \end{aligned} \tag{14}$$

where $x_1, x_2, x_3 \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.25 \\ 0.5 & 1 & 0.05 \\ 0.25 & 0.05 & 1 \end{bmatrix} \right)$, ATE $\tau = 0.25$, $\beta_1 = \beta_2 = \beta_3 = 1$, $\alpha = 0$ and $\varepsilon \sim N(0, 1)$. y_t represents the outcome variables for the treatment ($t = 1$) and control ($t = 0$) groups. Treatment is randomly assigned using a binomial distribution with probability $p = 0.4$.

5.1.1 Outliers

The contamination is added to the sample through replacing a small percentage of the data with a point mass outlier. Contamination is randomly determined with a binomial distribution with probability of $c = 0.01$. The contamination level is kept low, to keep this thesis' focus on infinitesimal levels of contamination. Outliers are put in the treatment group, the control group and in both groups together. The next paragraphs describe the exact contamination for the three different types of outliers: bad leverage points, good leverage points and vertical outliers. In all three cases, the contamination is such that it is hard to detect in simple explanatory analysis. This is done by plotting the outliers approximately 2 standard deviations from the variable's mean. This requires the computation of y's standard deviation:

As y is the sum of 4 normal distributions, y also follows a normal distribution. Its mean is obtained by simply adding all other distributions' means, which equals $\mu(0) = 0$ (control group) or $\mu(1) = 0.25$ (treatment group). For its variance, the following formula is used, provided that the x-variables are correlated to each other and ε is sampled independently.

$$\begin{aligned} Var(y) = {} & \beta_1^2 Var(x_1) + \beta_2^2 Var(x_2) + \beta_3^2 Var(x_3) + \\ & 2\beta_1\beta_2 Cov(x_1, x_2) + 2\beta_1\beta_3 Cov(x_1, x_3) + 2\beta_2\beta_3 Cov(x_2, x_3) + Var(\varepsilon). \end{aligned}$$

Filling in the numbers as presented in the DGP model (14), this gives $y_t \sim N\big(\mu_t, 5.525\big)$, with translates to an approximate standard deviation of 2.35.

Following the outlier classification in Section 3.2, the data manipulations to mimic outliers is described next. In the outliers formulas that follow, a * indicates a contaminated variable, meaning that variable lies far away from its mean. In this section, the contamination does not discriminate between treatment or control groups, but follows the random observations chosen by the binomial distribution over all observations. Note that the simulation also focuses on the scenarios where the contamination lies in the treatment or control group solely. The subscript t indicates the location of the outlier, either the treatment ($t = 1$) or control group ($t = 0$).

Firstly, bad leverage points are described: in this scenario, an observation's explanatory variables are altered, such that the point lies far away from the regression line. The outcome variables lies at its mean and the explanatory variables lie approximately two standard deviations from the mean.

$$z_{i,blp} = \Big[y_i = \bar{y}_t, x_{1i}^* = x_{2i}^* = x_{3i}^* = 2\Big] \tag{15}$$

Alternatively, the deviation from the regression line can be increased through changing the y-value in the opposite direction. A second type of bad leverage points is constructed as follows:

$$z_{i,blp2} = \Big[y_i^* = \bar{y}_t - 4.7, x_{1i}^* = x_{2i}^* = x_{3i}^* = 2\Big] \tag{16}$$

Secondly, good leverage points show deviating values for both its explanatory variables and outcome variable, but in line with the regression line for the uncontaminated sample. Good leverage points are constructed as follows:

$$z_{i,glp} = \Big[y_i^* = \bar{y}_t + 6; x_{1i}^* = x_{2i}^* = x_{3i}^* = 2\Big]. \tag{17}$$

Lastly, vertical outliers are also computed. Now, the outcome variable y is altered, such that the outcome variable is located at approximately two standard deviations from its mean, while the explanatory variables are in the middle of their distribution. Vertical outliers are constructed as follows:

$$z_{i,vert} = \Big[y_i^* = \bar{y}_t + 4.7, x_{1i} = x_{2i} = x_{3i} = 0\Big] \tag{18}$$

To illustrate the effect of outliers in the sample, a visualisation of the three types of outliers in both the treatment and control group is presented in Figure 2. The plots show the residuals against the squared Mahalanobis distance. The residuals are computed using an MM-estimator (Yohai, 1987), using the default settings of the *lmrob* function of package *robustbase* (Maechler et al., 2023) in *R* (i.e. Tukey-biweight loss functions for both steps, with respective tuning parameters $c_0 = 1.548$ and $c_1 = 4.685$) to provide a 50% breakdown point at a 95% efficiency level. For the (standardised) residuals, a horizontal cut-off point of 2.5 is chosen. The (squared) Mahalanobis distances are also robustly computed using the fast implementation of the MCD-estimator (Rousseeuw and Driessen, 1999). The vertical cut-off line is placed at $\chi^2_3(0.95)$. Note that these images represent a single sample and that the simulation study averages the results over 500 samples.

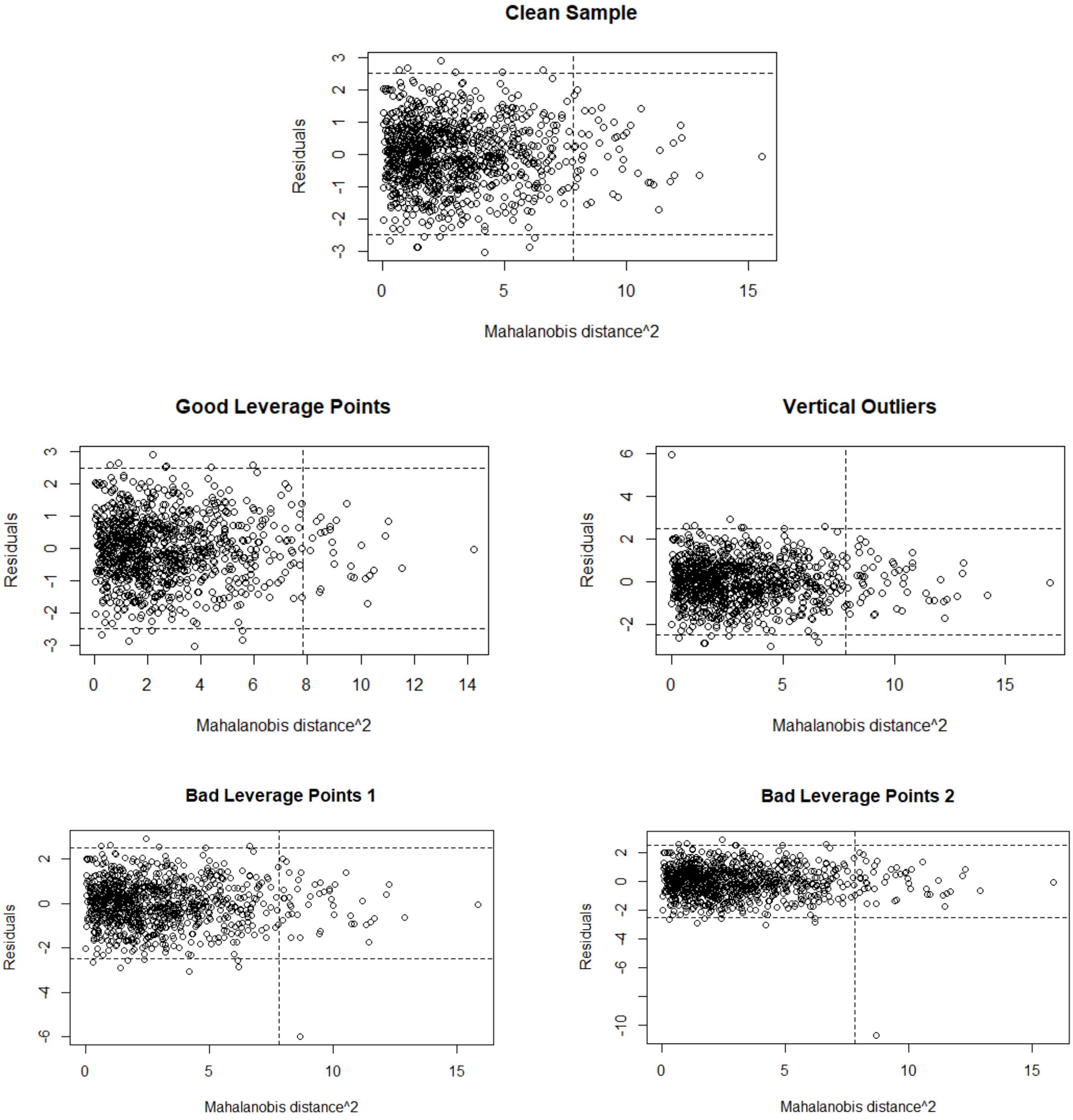


Figure 2: Exemplary outlier simulation.

5.2 National Supported Work Demonstration

For the real-data application, data from the National Supported Work (NSW) Demonstration is used. The NSW was a temporary employment program, operating in ten different sites across the United States in the mid-1970s. The program was designed to help disadvantaged workers get back into the working force by providing them with sheltered-environment jobs and counseling sessions. The duration of the program ranged between 9 and 18 months, afterwhich the workers were forced to find regular jobs. Unlike many other federal sponsored programs, the NSW program assigned qualified applicants completely randomly. Qualified applicants are AFDC women, ex-drug addicts, ex-criminal offenders and high-school dropouts. The initial study was conducted by LaLonde (1986), comparing both the NSW experimental setting with observational findings from different studies (i.e. Panel Study of Income Dynamics (PSID) and Current Population Survey-Social Security Administration (CPS-SSA) groups). LaLonde received much response to his work: e.g. Dehejia and Wahba (1999, 2002) introduced propensity score matching to the study and applied it to only a subset of the data, to include more variables in their analysis; Smith and Todd (2005) and Diamond and Sekhon (2013) further extended their work by comparing different matching and analysis techniques.

Following LaLonde (1986), this thesis specifically uses the male *experimental* NSW data, which is available in the *nws R* data package. In this sample, there are 297 treatment observations and 425 control observations (i.e. 41% of observations belongs to the treatment group). Applicant information available are his age, years of education, racial information (dummies for black and hispanic), marital status, high school dropout status, earnings in 1975 and earnings 1978. The latter is the outcome variable and all other variables are gathered pre-treatment. Data gathering was done through interviews, making earnings self-reported. In line with LaLonde (1986), all earnings are reported in 1982 USD. Simple descriptive statistics are reported in Table 2, where they are split by treatment group. Generally speaking, both groups are approximately similar and both strongly represent minority groups (80% is black, 11% is hispanic, and only 16% is married), are poorly educated (78% has no high school degree) and have low incomes (median income of $936.31, compared to a nationwide median income of $24,832.94[1]; 40% has no pre-treatment income at all). Moreover, the sample is relatively young, with an average age of 24.5 years old.

[1] $11,800 is the median income in 1975 (United States Consensus Bureau), but when corrected for inflation to get the equivalent in 1982 USD, the 1975 median income in 1982 USD is $24,832.94 (CPI Inflation Calculator). As a control measure: the reported median income in 1982 was $23,430 (United States Consensus Bureau)

Table 2: Descriptive statistics experimental (male) NSW sample.

Variables	**Treatment**	**Control**
Age	24.63 (6.686)	24.45 (6.590)
Education	10.38 (1.818)	10.19 (1.619)
Black	0.800 (0.400)	0.800 (0.400)
Hispanic	0.094 (0.293)	0.113 (0.317)
Married	0.168 (0.375)	0.158 (0.365)
No degree	0.731 (0.444)	0.814 (0.389)
Earnings 1975	3,066 (4,875)	3,027 (5,201)
Proportion 1975 zero-earners	0.374 (0.485)	0.419 (0.494)
Earnings 1978	5,976 (6,924)	5,090 (5,718)
Proportion 1978 zero-earners	0.226 (0.419)	0.304 (0.460)
Number of observations	297	425

The table values represent the sample group means, followed by their standard deviation in parentheses.

6 Methodology

This section covers the required methodology for each of the three treatment estimators. For each of them, both the classical estimators and their robust alternatives are discussed. The described methodology is applied using the aforementioned data with programming language *R* (R Core Team, 2023). Where needed, specific *R*-functions are mentioned to allow for reproduction. The seed is set to 123.

6.1 Difference-in-means estimators & robust alternatives

This section discusses the methodology for the difference-in-means estimator and its two robust alternatives. There are two separate sections for the value and standard error estimation, both also including details on implementation in *R*.

6.1.1 Estimators

Difference-in-means

The difference-in-means estimator is the first estimator to consider. It estimates the difference in the treatment group's outcome variable mean and the control group's outcome variable mean, via the arithmetic means. More specifically, this is computed as follows:

$$\hat{\tau}_{DiM} = \frac{1}{N_1}\sum_{i \in T}^{N_1} y_i - \frac{1}{N_0}\sum_{i \in C}^{N_0} y_i = \bar{y}_1 - \bar{y}_0.$$

Difference-in-trimmed-means
As shown in Section 4.2, the arithmetic mean is not a robust estimator for the population mean. One alternative that is used in this analysis is the difference-in-α-trimmed-means, where the $\alpha\%$ most outer observations are removed from the sample, prior to computing the arithmetic mean:

$$\hat{\tau}_{DiMT} = \frac{1}{N_1^*}\sum_{i\in T}^{N_1^*} y_i^* - \frac{1}{N_0^*}\sum_{i\in C}^{N_0^*} y_i^* = \bar{y}_1^* - \bar{y}_0^*,$$

where N_t^* represent the number of observations remaining in the treatment ($t = 1$) and control ($t = 0$) groups after trimming (both groups are α-trimmed separately). For each group, this is represented by $N_t^* = (1-\alpha)N_t$, such that $\frac{1}{2}\cdot\alpha\cdot N_t$ observations are trimmed from both the high- and low end of y_t. y_t^* is also the remaining subset of y's after trimming for both the treatment and control group. An advantage of trimming is the excluding of outliers in the estimation, therefore reducing the bias. On the downside, trimming too much can lead to overoptimistic conclusions due to estimating a too narrow standard deviation. Trimming can also be tricky for non-symmetric distributions. These two estimators are computed using the *difference-in-means* function from the *estimatr* package (Blair et al., 2022) in *R*. This research applies $\alpha = 0.1$, such that 5% of observations is removed on each end.

Difference-in-medians
Another robust alternative is the difference-in-medians estimator:

$$\hat{\tau}_{DiMed} = y_1^{(N_1+1)/2} - y_0^{(N_0+1)/2},$$

where the observations in both groups are separately ordered from lowest to highest and the estimate directly corresponds to the difference in *median* observations. In the case of an even number of observations, the estimate is the average between the middle two observations. Each of the two terms can then individually be replaced by $\frac{y_t^{N_t/2}+y_t^{(N_t/2+1)}}{2}$. The estimator is computed by using the *median* function twice, and evaluating their difference.

6.1.2 Standard errors

Difference-in-(trimmed-)means
For the difference-in-(trimmed-)means, standard errors are computed using the formula:

$$SE(\hat{\tau}_{DiM(T)}) = \frac{\bar{y}_1^{(*)} - \bar{y}_0^{(*)}}{\sqrt{\frac{(N_1^{(*)}-1)S_1^{2(*)}+(N_0^{(*)}-1)S_2^{2(*)}}{N_1^{(*)}+N_0^{(*)}-2}\left(\frac{1}{N_1^{(*)}}+\frac{1}{N_0^{(*)}}\right)}}, \tag{19}$$

where S_t^2 refers to the variance of y_t of the treatment ($t = 1$) or control ($t = 0$) group and * refers to the trimmed variants. The difference-in-means standard error is computed simultaneously with the value estimate by the *difference_ in_ means* function.

Difference-in-medians
Obtaining the standard error for the difference-in-medians estimator, requires a more manual approach: In the simulation study, the difference-in-medians estimate is computed $R = 500$ times, such that the standard error can be estimated from the sampling

distribution formed by the 500 estimates. For the NSW application, the data is not simulated repeatedly as is done in the simulation study. The difference-in-medians estimator (value) is computed using the *median* function, which only outputs a value estimate. For the NSW application, standard errors are computed using bootstrapping, where the difference-in-medians estimate is re-estimated B times. The B obtained estimates approximate the sampling distribution, therefore allowing standard error approximation. The bootstrapping procedure translates to repeating the following two steps B times:

1. The NSW data is resampled with replacement to obtain bootstrap sample b of size N.
2. For each bootstrap sample b, $\hat{\tau}_{DiMed,b}$ is computed (i.e. $\hat{\tau}_b$).

Then, the standard error can be approximated by:

$$SE(\hat{\tau}_{DiMed}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\tau}_b - \bar{\tau})^2},$$

where average $\bar{\tau} = \frac{1}{B}\sum_{b=1}^{B}\hat{\tau}_b$ and this thesis samples $B = 2000$ bootstraps in total. A similar procedure is applied to the difference-in-intercept estimators.

6.2 Regression-adjusted estimators & robust alternatives

This section discusses the methodology for the regression-based estimators described in Sections 4.3 and 4.4, meaning both the one-stage regression-adjusted estimator and the two-stage difference-in-intercepts estimator. In the classical variants, they are estimated using OLS. For the robust alternatives, an MM-estimator with Tukey beweight loss functions and the KW-estimator are implemented. All variants' methodology is presented for the two regression-adjusted estimators, including a description for implementation algorithms. Lastly, a brief paragraph is dedicated to describe the standard errors computation.

6.2.1 Estimators

OLS

The first estimator to be discussed is the classical OLS estimator. Even though the OLS estimator is an M-estimator (Huber, 1964) with a squared loss function, $\hat{\beta}$ can also be estimated by

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y,$$

Note that notation for X here is in capitals, referring to the whole data set, instead of individual level, as was the case in Section 4 (i.e. instead of p-length vector x, here $n \times p$ matrix X is used). For both regression-adjusted estimators, this is solved through using *R* function *lm* from the *stats* package.

MM-estimator

As a robust alternative to OLS, MM-estimators are implemented. As already mentioned, MM-estimators combine M-estimators and S-estimators. The MM-estimator is a multi-stage estimator. Prior to evaluating the stages, the M- and S-estimator elements are presented: For M-estimators (Huber, 1964) of *location*, $\hat{\beta}_M$ is

$$\hat{\beta}_M = \underset{b}{\operatorname{argmin}} \sum_{i=1}^{N} \rho\Big(\frac{y_i - x_i^T b}{\sigma}\Big).$$

and M-estimators of *scale* $\hat{\sigma}_M$ can be found as the solution for

$$\frac{1}{N}\sum_{i=1}^{N} \rho\Big(\frac{y_i - x_i^T b}{\hat{\sigma}_M}\Big) = \delta, \tag{20}$$

for a given value b (i.e. $\hat{\sigma}_M(b)$), where $\delta = E_F\Big[\rho(\frac{x}{\sigma})\Big]$ (for consistency at model distribution F) and ρ is a chosen loss function, validating some regulatory assumptions. Next, S-estimators (Rousseeuw and Yohai, 1984) of location can be computed using the M-estimator of scale σ_M:

$$\hat{\beta}_S = \underset{b}{\operatorname{argmin}}\, \hat{\sigma}_M^2(b), \tag{21}$$

where $\hat{\sigma}_M(b)$ is the solution of equation (20). In turn, the S-estimator of scale equals $\hat{\sigma}_S = \hat{\sigma}_M(\hat{\beta}_S)$. It is evident that the S-estimator is dependent on the M-estimator for both its location and scale estimates. The M-estimator of scale is dependent on a b estimate, for which $\hat{\beta}_S$ turns out to be a good fit. Also, location S-estimator equation (21) can be rewritten as

$$\hat{\beta}_S = \underset{b}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\Big(\frac{y_i - x_i^T b}{\hat{\sigma}_S}\Big). \tag{22}$$

Now that all sub-parts and underlying mechanisms are established, the MM-estimator steps can be provided (Yohai, 1987):

1. Obtain initial scale location estimate T_0:
   It is critical for the breakdown point of the final estimator that this initial estimate has a high breakdown point. In this thesis, an initial S-estimator is used.

2. Compute residuals & M-scale estimate:
   The residuals $r_i(T_0) = y_i - T_0^T x_i$ are computed and used as input for computing the M-scale $\hat{\sigma}_M(r_i(T_0))$. In this step, the loss function ρ_0 in equation (20) for solving $\hat{\sigma}_M(r_i(T_0))$ is the Tukey biweight loss function with tuning parameter $c_0 = 1.548$, tuned for high breakdown.

3. Compute the MM-estimate $\hat{\beta}_{MM}$:
   $\hat{\beta}_{MM}$ can be found through the location S-estimator:

   $$\hat{\beta}_{MM} = \underset{b}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_1\Big(r_i(b)/\hat{\sigma}_S\Big), \tag{23}$$

   where $\hat{\sigma}_S = \hat{\sigma}_M(r_i(T_0))$ and ρ_1 is the second Tukey biweight loss function, with tuning parameter $c_1 = 4.685$, tuned for high efficiency.

The estimator is estimated using the *lmrob* function from the *robustbase* package (Maechler et al., 2023), which is further explained in the next section.

KW-estimator
Krasker and Welsch (1982) names the estimator the modified least squares (MLS) estimate, which is defined by $\hat{\beta}$ in

$$\sum_{i=1}^{N} \Psi\big\{(y_i - x_i\hat{\beta})A^{-1}x_i^T\big\} = 0,$$

where Ψ is the Huber score function and matrix A is a $k \times k$ matrix satifying

$$A = E\Big[2\phi(\frac{a}{\sigma|A^{-1}x^T|}) - 1\Big]x^T x.$$

Matrix A exists for sufficiently large a, specifically

$$a \geq \sqrt{\frac{\pi}{2}}\frac{\sigma}{E||x||}.$$

a represents the tuning parameter of the Huber function. The Huber loss- and score functions are as follows:

$$\rho(r) = \begin{cases} \frac{r^2}{2}, & \text{if } |r| \leq a \\ a(|r| - \frac{a}{2}), & \text{otherwise} \end{cases}$$

$$\Psi(r) = \begin{cases} r, & \text{if } |r| \leq a \\ a \cdot sign(r), & \text{otherwise} \end{cases}$$

Note that the Huber score (loss) function equals the least squares score (loss) function for $|r| \leq a$.

6.2.2 Solving algorithms

OLS
The OLS estimates can be constructed in one simple step, as the formula provided in Section 6.2.1 only includes the (raw) data and does not other required estimates. No additional algorithms are necessary, beyond the aforementioned *lm* function.

MM-estimator
MM-estimates are in theory more complex and require multiple solving algorithms. MM-estimates are constructed using the *lmrob* function from the *robustbase* package (Maechler et al., 2023). As an alternative to the "*MM*" method, this function's method can also be called the "*SM*" method, because it computes an S-estimate as the starting value for the M-estimate afterwards. This thesis refers to the two stages as the S-step and the M-step. For each of the steps, the *lmrob* function performs a different algorithm, which is discussed in the next paragraphs.

In the S-step, it is important that the starting values have a high breakdown point. Since S-estimators can be tuned for a high breakdown point, it is an effective choice to use an S-estimator. However, unlike in the the third step by Yohai (1987), the computation for the *initial* estimator cannot be dependent on another (scale) estimate. Treating the S-estimator as a weighted least squares estimator, Salibian-Barrera and Yohai (2006) developed the FAST-S estimator (available in *R* in the *lmrol.S* function in the *robustbase* package (Maechler et al., 2023)) based on so-called improvement steps. These improvement steps are repeated until convergence and form the middle section of the I-step algorithm:

1. Start with m sets of starting values β_0
2. For all m sets, perform the improvement steps until convergence
3. Return $\hat{\beta}_S$ corresponding to the estimator with the lowest M-scale estimate over all m sets,

where one improvement iteration (step 2) consists of the following three steps, beginning with current estimate $\hat{\beta}_k$:

1. Compute M-estimate of scale $\hat{\sigma}_M(\hat{\beta}_k)$
2. Update weights
$$w_i(\hat{\beta}_k) = \frac{\Psi_0\Big((y_i - x_i^T\hat{\beta}_k)/\hat{\sigma}_M(\hat{\beta}_k)\Big)}{(y_i - x_i^T\hat{\beta}_k)/\hat{\sigma}_M(\hat{\beta}_k)}$$
3. Obtain a new location estimate $\hat{\beta}_{k+1}$
$$\hat{\beta}_{k+1} = \underset{b}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} w_i(\hat{\beta}_k)(y_i - x_i^T b)^2$$

Then, for every iteration $\hat{\sigma}_M(\hat{\beta}_{k+1}) \leq \hat{\sigma}_M(\hat{\beta}_k)$. In this thesis, Ψ_0 used in step 2 represents the Tukey biweight score function with tuning parameter $c_0 = 1.548$ (the default value in *R*).

Now that the starting value $\hat{\beta}_S$ and corresponding scale estimate $\hat{\sigma}_S = \hat{\sigma}_M(\hat{\beta}_S)$ are established, the function continues with the M-step. In the M-step, an IRLS algorithm is run, which resembles the following, for current estimate $\hat{\beta}_k$:

1. Update weights
$$w_i(\hat{\beta}_k, \hat{\sigma}_S) = \frac{\Psi_1\Big((y_i - x_i^T\hat{\beta}_k)/\hat{\sigma}_S\Big)}{(y_i - x_i^T\hat{\beta}_k)/\hat{\sigma}_S}$$
2. Obtain weighted least squares solution
$$\hat{\beta}_{k+1} = \underset{b}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} w_i(\hat{\beta}_k, \hat{\sigma}_S)(y_i - x_i^T b)^2,$$

where it is important to note that $\hat{\sigma}_S$ does not change throughout the iterations and that Ψ_1 used in step 1 represents the Tukey biweight score function with tuning parameter $c_1 = 4.685$ (the default value in R). The algorithm starts with $\hat{\beta}_0 = \hat{\beta}_S$ and is repeated until convergence. Evidently, the finding of $\hat{\beta}_{MM}$ is dependent on $\hat{\sigma}_S$, which affects which observations are excluded (i.e. given a low weight) from the analysis, i.e. filtering out the outliers. This is expected to work nicely for vertical outliers and bad leverage points, but as good leverage points actually contribute to the decrease of $\hat{\sigma}$, it is expected that good leverage points stay included in the MM-regression.

KW-estimator

The KW-estimator is computed following the SAS™ software code in Mehta (2023). The algorithm consists of three steps:

1. Compute OLS start values and compute the DFFITS values for each observation. The DFFITS-value for an observation is the difference in OLS estimates with or without omitting that observation:

   $$d(x_i) = DFFITS_i = \frac{x_i\hat{\beta} - x_i\hat{\beta}(i)}{s(i)\sqrt{h_i}} = \frac{h_i^{\frac{1}{2}}(y_i - x_i\hat{\beta})}{s(i)(1-h_i)} \text{ (Krasker and Welsch, 1982),}$$
   $$= (x_i A^{-1} x_i^T)^{\frac{1}{2}}$$

   such that $A = \frac{1}{N}\sum_i w_i^2\{d(x_i)\}x_i^T x_i$. $s(i)$ and $\hat{\beta}(i)$ are the usual estimates for σ and β, but without the ith observation and $h_i = x_i(X^TX)^{-1}x_i^T$ (i.e. the diagonal elements of hat matrix $X(X^TX)^{-1}X^T$). The residuals of this regression are called RES_{old}.

2. Define new variables and compute weights for a weighted OLS estimate:

   $k = 1.5\sqrt{p} \cdot \sqrt{\frac{p}{N}}$, setting parameter a to $1.5\sqrt{p}$, as suggested by Krasker & Welsch

   $$kwl = \frac{k}{|DFFITS|}$$
   $$w = \begin{cases} kwl, & \text{if } kwl \leq 1 \\ 1, & \text{otherwise} \end{cases}$$

   Compute a new estimate using the computed weights and, again, compute the DFFITS-values. ($DFFITS_{new}$). The residuals of this regression are called RES_{new}.

3. Update the weights using $DFFITS_{new}$, now adjusted by the previous residuals:

   $$DFFITS_i = DFFITS_{new} \cdot \frac{RES_{new}}{RES_{old}}$$

   and re-iterate the weighted OLS estimate. Set $RES_{old} = RES_{new}$ and compute new values for $DFFITS_{new}$ and RES_{new} using the latest weighted regression model.

Step 3 is iterated until convergence (i.e. the maximum difference between the current and new estimate is 0.05) or until a number of iterations is reached (for this thesis, this is set to 500). The algorithm uses the *lm* function to run the (weighted) regressions and uses the *DFFITS* function from the *CRAN* package to compute the DFFITS values.

6.2.3 Standard errors

OLS
For the OLS estimators, robust standard errors are computed using the packages *sandwich* and *lmtest* in *R*. This is done to correct reported standard errors for heteroskedasticity in the sample (i.e. heteroskedasticity-consistent standard errors). Classic standard errors are replaced with Huber-White standard errors (White (1980a, 1980b)), giving variance $(X^TX)^{-1}X^T\Omega X(X^TX)^{-1}$. Ω is estimated by sample covariance matrix S (classical OLS variance estimation uses $\Omega = \sigma^2 I$, such that the variance reduces to $\sigma(X^TX)^{-1}$). The *R*-function *vcovHC* in package *sandwich* (Zeileis et al., 2022) allows for different estimation methods, providing more flexibility to the OLS assumption of constant variance. This thesis uses the "*HC0*" type, such that $S = diag(e_1^2, ..., e_N^2)$.

MM-estimator
For the MM-estimators, the asymptotic variance formula as mentioned in Croux et al. (2004) is used. The paper mentions multiple variants, but this thesis uses "Avar$_{1s}$", which is robust against outliers and heteroskedasticity. However, it does assume symmetric error terms, which may be too restrictive and possibly does not hold in practice. The asymptotic variance is computed as follows:

$$\begin{aligned} Avar_{1s}(\hat{\beta}_{MM}) &= AE(\Psi^2 X^T X)A \\ &= \sigma^2[E(\Psi X^T X)]^{-1}E(\Psi^2 X^T X)[E(\Psi X^T X)]^{-1}, \end{aligned} \tag{24}$$

where $A = \sigma[E(\Psi^T X^T X)]^{-1}$. The regression coefficient standard errors for covariate j can then be computed through:

$$\hat{se}(\hat{\beta}_{MM,j}) = \sqrt{\frac{1}{N}\widehat{Avar_{1s}}(\hat{\beta}_{MM})_{jj}}$$

for $j = 1, ..., p$. When estimating the elements, $\hat{\beta}_{MM}$ and $\hat{\sigma}_S$ are used (also for $\hat{\Psi}$) and expectation $E(.)$ is approached by $\frac{1}{N}\sum_{i=1}^{N}$.

KW-estimator
The KW-estimator is a weighted-OLS estimate. Therefore, the (HC0) standard errors are similar as for the regular OLS estimator, but now including the weights, such that the variance is $(X^TWX)^{-1}X^TW\Omega WX(X^TWX)^{-1}$, where $W = diag(w_1, ..., w_N)$.
bigbreak

7 Results

7.1 Simulation

This section presents the simulation results based on the data generation process and outlier contamination as described in Section 5.1. The results are illustrated in Figures 4 and 5, grouped by type of contamination, but an alternative representation (grouped by estimator) can be found in Appendices B.2 and B.3 (a de-medianed version to better allow for comparison in changes in variance). As a baseline, the estimation results are presented for the clean sample in Figure 3 below. This section only presents and comments on the treatment effect estimates, but estimates for the full set of covariates can be found in Appendix B.1.

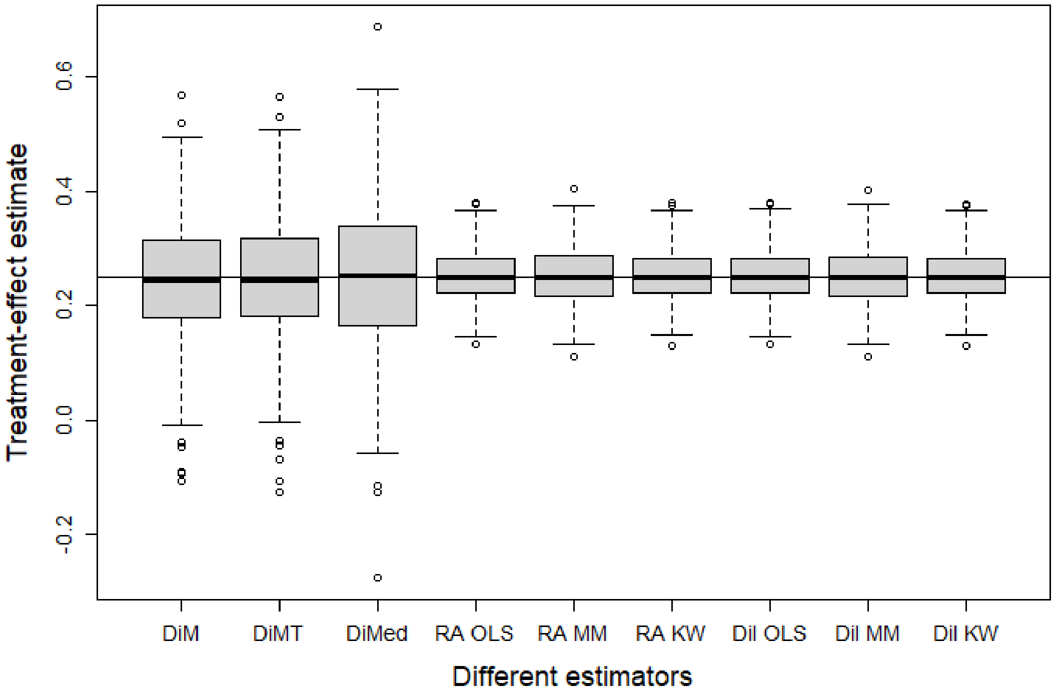
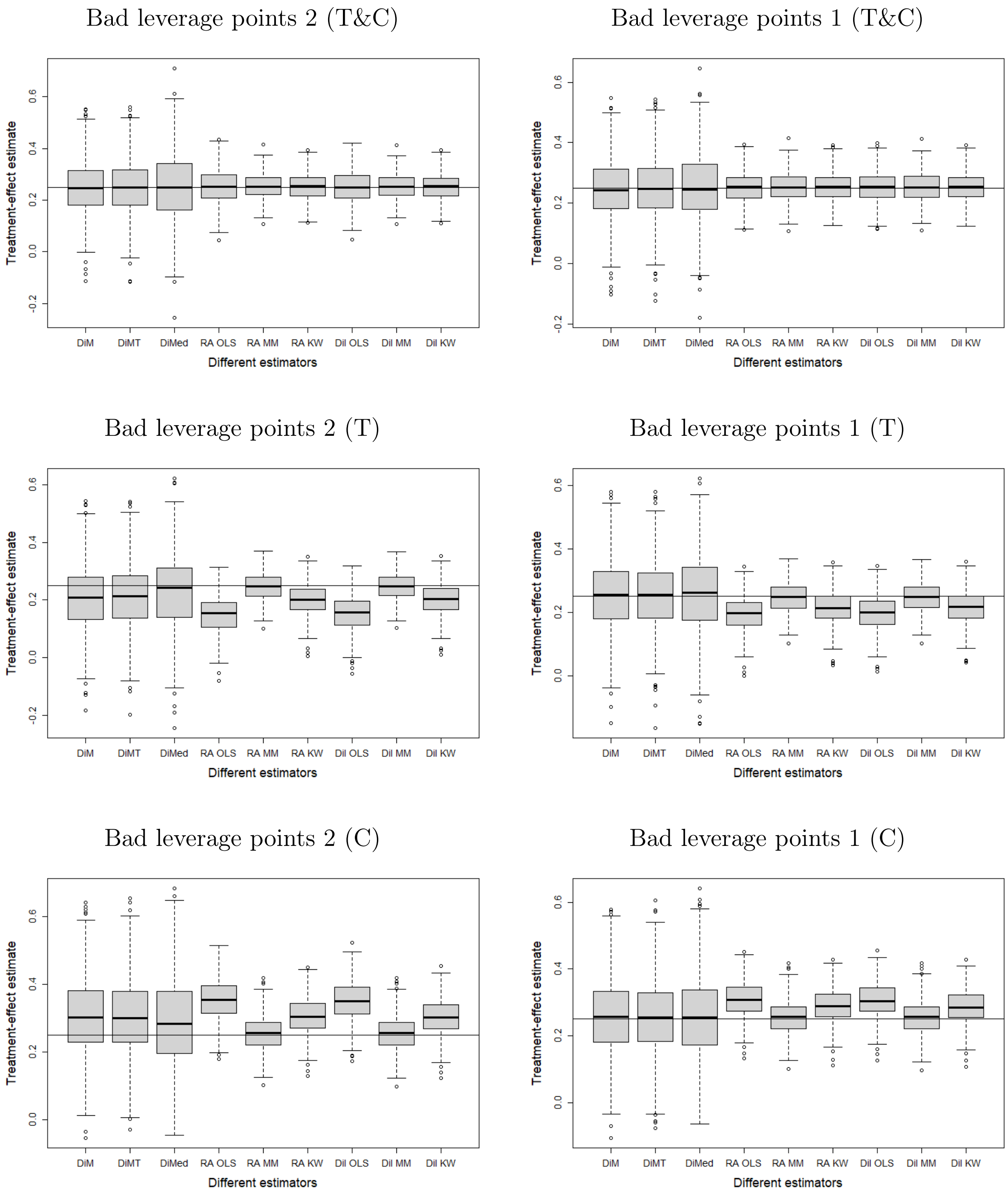


Figure 3: Estimation results for the clean sample

As expected, all estimators are consistent in the clean sample and there is a clear difference in variance among the estimators. The difference-in-means based estimators show significantly larger variances compared to the estimators including covariates in the analysis. Furthermore it shows that the classical estimators have a lower variance and that between the regression-adjusted and difference-in-intercepts estimators, the variances are equal.

After having evaluated the estimators in the clean sample, assessing the effect of outliers is next. Boxplots for the different estimators for each type of contamination are presented in Figures 4 and 5. Firstly, findings corresponding to the difference-in-means based estimators are presented, followed by a discussion on the covariates-including estimators. Lastly, the two groups are compared and contrasted, with a focus on the differences between the "classical" estimators (i.e. the difference-in-means and the regression-adjusted estimator using OLS). For variance inferences, centered boxplots are created for an easier comparison between the estimators (Appendix B.3).



Bad leverage points 2 (T&C)

Bad leverage points 1 (T&C)

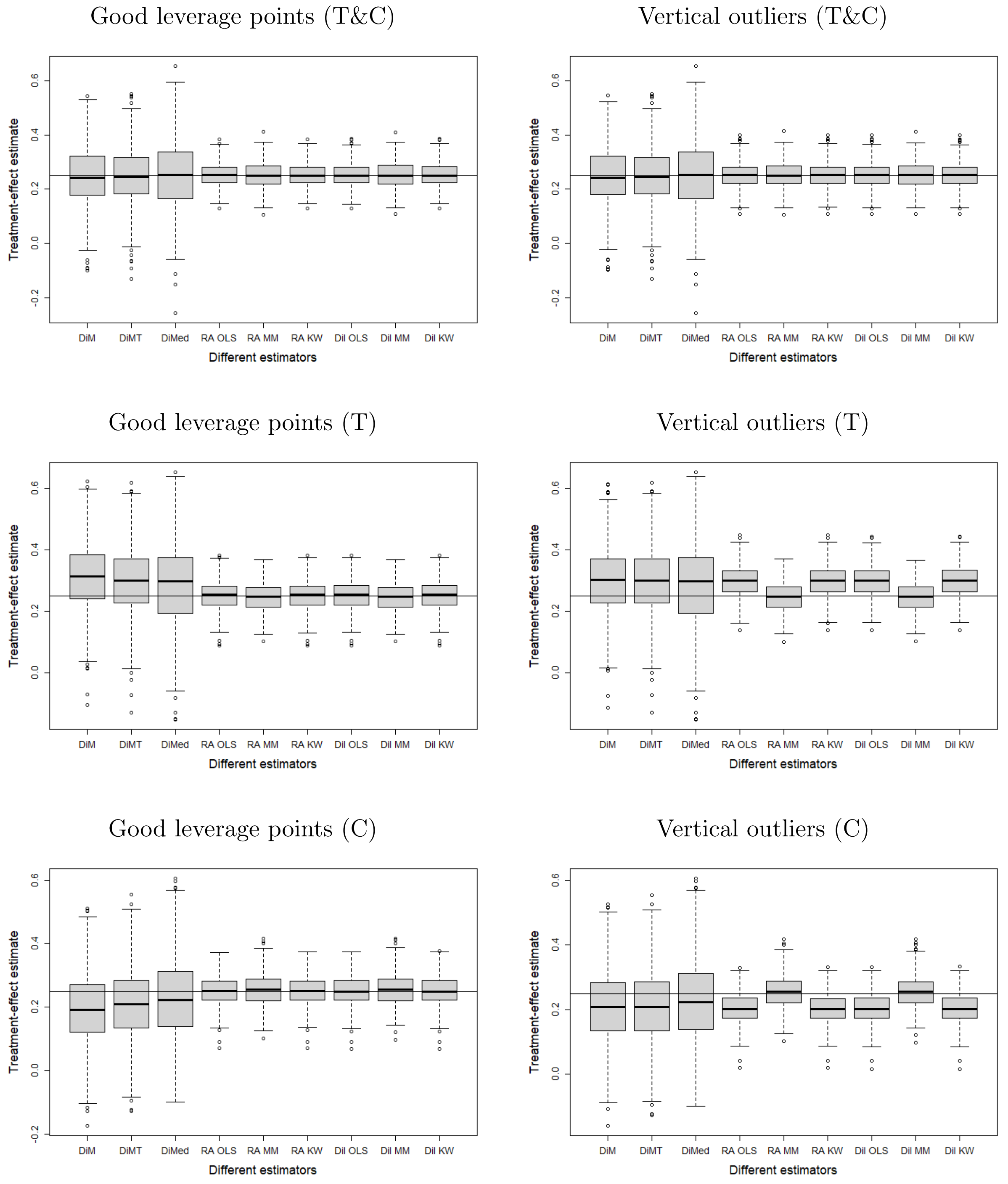
Bad leverage points 2 (T)

Bad leverage points 1 (T)

Bad leverage points 2 (C)

Bad leverage points 1 (C)

Figure 4: Boxplots treatment effect estimation bad leverage points


Good leverage points (T&C)

Vertical outliers (T&C)

Good leverage points (T)

Vertical outliers (T)

Good leverage points (C)

Vertical outliers (C)

Figure 5: Boxplots treatment effect estimation good leverage points and vertical outliers

Firstly, a comparison between the difference-in-means estimators is made: Across all forms of contamination, the three estimators react in the same direction, but the difference-in-medians estimator often stays closest to the true treatment effect. For all three estimators, the estimates are biased for the bad leverage points 2, good leverage points and vertical outliers, when the contamination is present in only one of the groups. Moreover, the differences between the difference-in-means estimator and its trimmed variant are minimal, because the outliers, as simulated in this study, do not lie far enough in the tails to be trimmed away (only noticeable differences for good leverage points, whose

y-values lie more than 2 standard deviations away from its mean). Comparing variances, it can be noted that the difference-in-medians estimator has a larger absolute variance, but that it is less volatile to outliers. Moreover, when the center of the distribution thickens (as is the case with bad leverage points 1), the difference-in-medians' variance decreases noticeably, while the other two estimators are more invariant to observation distributions (provided similar means).

Secondly, the regression-adjusted and the difference-in-intercepts estimators are discussed: It is very evident that the OLS-estimators are a lot more unstable than the MM- and KW-estimators (they are in fact the most volatile estimators in this study). More specifically, the OLS-estimates are strongly biased for bad leverage points and vertical outliers. For good leverage points, however, they are consistent and have a decreased variance estimate (as expected). Furthermore, for vertical outliers in both groups and in the control group only, the OLS variance estimates are lower compared to that of the clean sample, whereas this is not the case for vertical outliers in the treatment group (see Appendix B.3 for a clearer comparison). This is probably because the variance (partly) "moves" towards the regression intercept estimate in the former two groups, whereas in the latter all variance remains for the treatment effect estimate. In contrast, the MM-estimators are very stable both in terms of bias and variance. Similarly to the OLS-estimator, the KW-estimator is also unstable and biased in the presence of outliers: For bad leverage points, its performance lies in between that of OLS and MM, but for vertical outliers it has not improved compared to OLS. Zooming in, this can be explained due to the KW-estimator identifying the bad leverage points well, but the weights are too far from 0 to remove all aberrant effects. For vertical outliers, the outliers are not identified correctly and no observations are signficantly down-weighted.

Lastly, differences between the two aforementioned groups of estimators is discussed. A first observation worth noting, is that for all four types of contamination, on average, all estimators stayed consistent when the contamination was evenly spread between both groups (i.e. 1% contamination in both groups/randomly over the entire sample). It is likely the case that the contamination in one group is canceled out by the other, as is visible when evaluating intercept estimates in Table 5 in Appendix B.1. However, the non-robust estimators' variances then did increase, relatively to the clean sample. Furthermore, similar to the clean sample case, estimator variances remained significantly reduced with the addition of covariates in the estimators, even when those covariates are contaminated. Note that this observations cannot automatically be extended to all variations of contamination, but refers to the contamination as simulated in this thesis. Lastly, comparing the "classical" difference-in-means estimator and the OLS-regression estimator, there are a few differences to be noted: In general, the OLS estimator reacts more strongly to outliers in term of bias in the value estimates and increases in relative variance. This holds specifically true for vertical outliers and bad leverage points 2. However, for all types of contamination, the regression-adjusted variances remained only a fraction of that of the difference-in-means estimate. Moreover, OLS-estimates are robust to good leverage points, while the difference-in-means estimator is not. In turn, OLS fails for bad leverage points 1, where the difference-in-means estimator remains unbiased there.

7.2 NSW data application

This section provides the results from using different estimation methods on the NSW data set. The results are summarised in Table 3 (estimates for all additional covariates can be found in Appendix C.1) and it must be noted that none of the regression-adjusted estimates passed the Shapiro-Wilk's test for normality of the error terms. The table includes the results from the three difference-in-means based estimators; a simple regression as presented in the data's original paper by LaLonde (1986), using the same covariates; the regression-adjusted estimator similar to the previous estimator, now with demeaned covariates in an interaction term with the treatment indicator as introduced by Lin (2013); and its two-step equivalent as mentioned in Lei and Ding (2020), using demeaned covariates as well. It must be noted, that the first two results in the first line replicate the results as presented in LaLonde (1986), but here "HC0" type standard errors are reported.

Table 3: NSW results treatment effect

Difference-in-means		**Linear Regression**		**Lin (2013)**		**Difference-in-intercepts**	
Means	886.30* (488.20)	OLS	791.44* (485.65)	OLS	785.37 (483.57)	OLS	785.37 (493.25)
Trimmed	718.38* (368.07)	MM	432.83 (401.31)	MM	412.62 (403.82)	MM	574.94 (761.86)
Medians	485.61 (648.82)	KW[1]	789.26 (484.77)	KW[1]	754.77 (472.05)	KW[1]	768.86 (478.42)

1: The KW-estimators did not converge, but bounced between two local minima. Estimates of one minima are still included because the estimates were relatively close to each other and to show the similarity to the OLS estimates.

Treatment-effect estimates, robust standard errors in parentheses.*** p<0.01, ** p<0.05, * p<0.1

It is immediately visible that there is much variability in treatment effect estimates between the estimators. This suggests the presence of outliers in the sample, therefore creating interest for investigating the sample in more detail. Figure 6 shows the observations' standardised residuals plotted against their squared Mahalanobis distances. For the standardised residuals, the MM-estimator as in Lin (2013) is used and the squared Mahalanobis distances are computed using the MCD-estimator (fast implementation as introduced by Rousseeuw and Driessen (1999)) center and covariance. The cut-off lines are placed at 2.5 and $\chi^2_7(0.95)$, respectively:

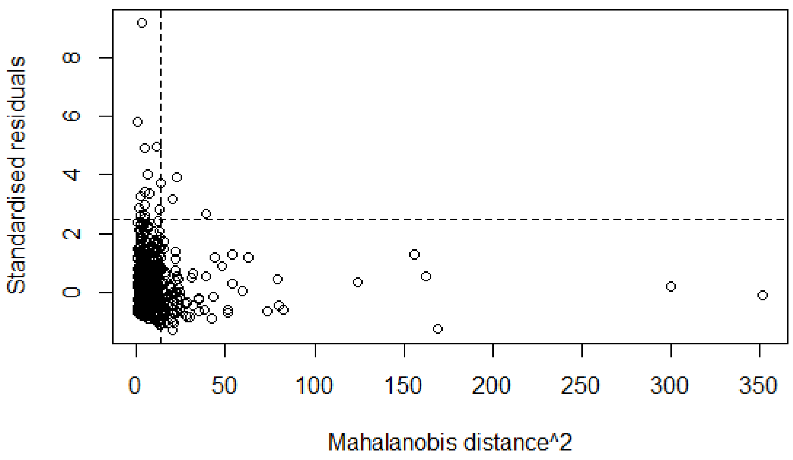


Figure 6: Robust standardised residuals against squared Mahalanobis distance

The plot shows data with many outliers, both in the explanatory variables and the outcome variables. Vertical outliers are observations with a small Mahalanobis distance, but a large residual (15 observations); Good leverage points have small residuals, but a large Mahalanobis distance from the center (90 observations); and bad leverage points are those observations in the top right box (3 observations). So, the majority of the outliers are good leverage points, followed by a fair share of vertical outliers. In total, 15% of the data is categorised as outliers, with 16,5% contamination in the treatment group and 13,9% in the control group. Similar plots, now excluding *agesquared* and plots for the two treatment groups individually can be found in Appendix C.2. When *agesquared* is omitted, the squared Mahalanobis distance "only" ranges till 65 and observations in minority groups (i.e. married, hispanic, caucasian, etc) are more present "at the top", rather than the Mahalanobis distance being predominantly age-dependent). Complementing the plots for the separate treatment groups, Tables 10 and 11 in Appendix C.3 and C.4, respectively, provide an overview of the data for the top 15 standardised residuals (vertical outliers & bad leverage points) and Mahalanobis distances (good leverage points), respectively. In the former table, the 13 observations with the highest standardised residuals are also the 13 highest earners (1978 income).

As a further deep-dive, it is interesting to evaluate the outcome variable's histogram (outliers in the y-direction). Figure 7 shows histograms of the outcome variable, for both the treatment and control groups. Both groups are strongly skewed to the right and have a long and narrow tail. The treatment group specifically has one observation whose outcome variable lies extremely far away from the majority of observations. For reference, removing this one observation from the treatment group, decreases its mean by \$76.08. Moreover, the difference-in-means estimate for earnings over $\$20,000$ (17 observations in total) is $\$4,774$, whereas this is only \$515 for the group with earnings below $\$20,000$ (it drops even to $\$-7$ when excluding the group with earnings $= 0$). This shows that the difference-in-means estimate is strongly affected by the relatively small group of men in the higher end of earnings (for the top 10 largest earnings, the difference-in-means estimator reaches $\$8,819$).

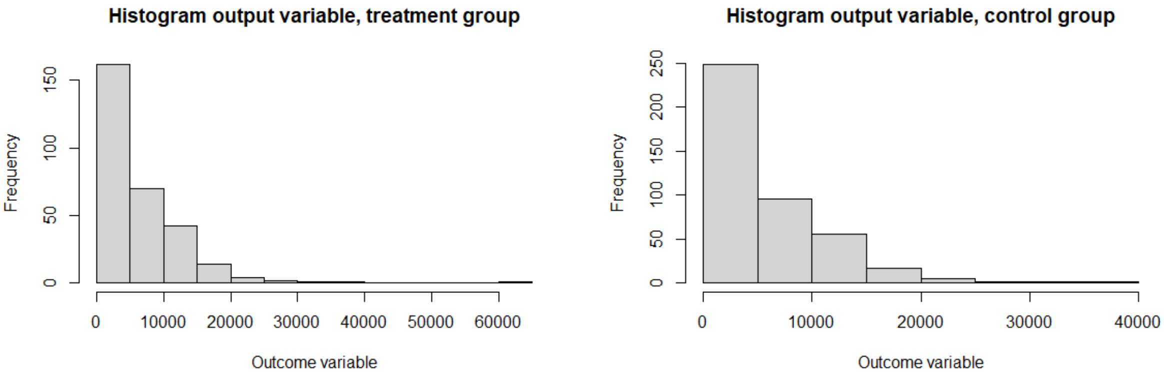


Figure 7: Histograms outcome variable (earnings 1978)

Next, it is interesting to evaluate the (differences in) estimates in Table 3. In the first column, it can be noted that the different estimators provide significantly different values for the treatment effect. This is attributable to the strong skewness in the outcome variable's distribution. For both groups, trimming on the left side removes only \$0-earners while trimming on the right removes the top 5% of earners, which represent most of the distribution's long tail, specifically for the treatment group. This differences in right tails, explain why the difference-in-trimmed-means estimate is lower than the regular difference-in-means estimate. It is likely that the trimmed variant therefore presents a better image of the true effect, but it is difficult to determine the best cut-off point and to know when the observations are outliers or not, and thus to set the trimming percentage correctly. The third estimator in this category is the difference-in-medians estimator. It is difficult to draw conclusions about the validity of this estimator. On the one hand, the median and mean do not coincide in non-symmetric distributions, but the sample median would then be smaller, therefore making the median no longer a consistent estimator for the mean. On the other hand, there are some obvious outliers in the sample (undoubtedly the largest earnings observation in the treatment group), making the median a better fit for centrality. Additionally, the difference-in-medians estimate lies significantly closer to the regression-adjusted MM-estimates than the other two do. Furthermore, considering their variances, the findings align with those of the simulation study, where the median variant has the largest variance and the trimmed-mean variant has the smallest.

Comparing the middle two columns (the regression-adjusted estimates), there is a clear division present between the different estimators, as is to be expected in a sample with outliers. The MM-estimator removes/significantly downweights (i.e. weights < 0.1) 15 observations, including the 13 highest earners from the analysis, which are the strongest vertical outliers and all three bad leverage points (see Appendix C.3 for details on those observations). After doing so, there is no longer enough evidence to conclude that the NSW employment program has a significant effect on the participants' posterior earnings. Despite not fully converging, the KW-estimates resemble more to the OLS-estimates than to the MM-estimates, for both variance and value estimates. This might be because of the larger share of vertical outliers in the sample (in the simulation, it can be noted that the KW-estimator performs similar to OLS in the presence of vertical outliers) or because the percentage of outliers in the sample is higher and the KW-estimator simply breaks down. In any case, it failed to identify the outliers correctly and the handful of observations that were downweighted, still had weights relatively close to 1. Moreover, the MM-estimates have the lowest standard errors, which again signals there are outliers in the sample. However, as established earlier, the majority of the outliers are good leverage points and none of the estimators filters those out. Therefore, it is likely that all three estimators underestimate the variance.

Furthermore, the differences between the simple linear regressions and the Lin (2013) estimators are small, signaling negligible heterogeneous effects of the covariates on the outcome variable, across the treatment and control groups. Moreover, none of the added interaction terms have a significant effect on earnings in 1978 (Appendix C.1). Also, roughly half of the covariates originally included in the analysis by LaLonde (1986) do not have a significant effect on the outcome variable, which can explain the (slight) increase in standard error in the MM-Lin (2013) estimator (i.e. adding the interaction terms leads to a decrease in degrees of freedom and lack of additional explanatory value). Probably attributable to the outliers, but the same does not hold true for OLS.

Lastly, the regression-adjusted estimators are compared to the difference-in-intercepts estimators. For OLS, the estimates are equivalent (there is a slight difference in standard error, which is due to the bootstrapping of the latter estimator's standard error), as was expected. Interestingly, this does not hold for the MM-estimators. Here, we see an estimate that deviates from the regression-adjusted MM-estimates, with a significantly enlarged standard error. Checking differences in weights gave no significant results: Most observations removed in the Lin (2013) regression are also removed in the two individual MM-regression for the difference-in-intercepts estimate, and vice versa. There were some small differences, but the weights remained small (i.e. weights < 0.15, in both MM-regressions). Another explanation can be with regards to the bootstrapping procedure to estimate the difference-in-intercepts' variance. Provided that there are many good leverage points, there could exist bootstrap samples with (relatively) many good leverage points (in only one treatment group), such that those outliers dominate the regression line. Hence, they can affect the intercept estimate, creating a wide array of estimates and therefore a large variance estimate. Similar to the regression-adjusted estimators, the KW-estimator did not converge but bounces between two local minima. Results for one of those minima is still included to show the resemblance to OLS.

8 Discussion and Conclusion

This thesis has researched the effects of infinitesimal level of contamination on three different treatment-effect estimators. For each classical estimator, the IF and CVF are derived and it can be concluded that all formulas are unbounded. Furthermore, the simulation and NSW data study also show interesting differences between the different estimators, which are discussed in this section.

The derivation of the IF and CVF functions for the classical estimators are in line with previous literature. Treatment effect estimators including the arithmetic mean or those which apply OLS are prone to outliers and have both an unbounded IF and CVF. This finding is also confirmed by the simulation study, where the estimates are biased and often have an inflated variance in the presence of outliers. Moreover, for small levels of contamination and non-extreme outliers (i.e. this thesis' simulation) there is a clear stability-variance trade-off between the regression-adjusted OLS and difference-in-means estimators. Though both estimators are biased, the OLS estimate reacts significantly more strongly to outliers than the difference-in-means estimator does, both in terms of bias and variance increase. That being stated, OLS variance as an absolute number still remains lower than that of the unadjusted estimator for all types of contamination. This relation is probably variable to the exact type of contamination and can differ for different settings of outliers (i.e. different level of contamination and location of point mass contamination).

Starting with the difference-in-means estimators, the robust alternatives under-performed in terms of estimate improvement. Despite the median being the most robust estimator of location in the literature (Hampel et al., 1986), its estimates had a visible bias in the presence of outliers (because of its local-shift sensitivity) and its variance is significantly larger than that of the mean. This shows the robust estimators do not necessarily provide unbiased results in practice. However, despite its larger variance in absolute terms, it did present a much more stable variance in the presence of outliers. Moreover, the relatively large bias can be explained by the fact that the treatment effect to be estimated is small, causing a small deviation to have a relatively large effect on the bias. This suggests that the difference-in-medians estimator performs better when the treatment effect is larger (i.e. the same distribution, only with a larger mean). Similarly, an increase in the number of sampled observations or a more dense distribution around the median can possible also decrease the bias, as the values of the near observations are more alike, and hence a smaller deviance is obtained in the presence of outliers. Furthermore, the difference-in-medians plots remain unchanged for far more extreme point mass contamination, as the median estimates are solely influenced by the observations around the initial median, which remain unchanged. The same holds true for the difference-in-trimmed-means estimator. In the case of more extreme point mass contamination, the estimator even improves, because it is now more successful in removing the outliers. Furthermore, the latter estimator has a significantly smaller variance compared to the difference-in-medians estimator. Again, there is an interesting trade-off between these two estimators: both can handle extreme outliers well, but the median is better resistant against larger levels of contamination, whereas the trimmed-mean is preferred for its smaller variance in scenarios with smaller levels of contamination. However, it is difficult to determine a good trimming percentage. Therefore, the median can be used as a first step to prioritise robustness and different trimming percentages can be used as a second step to possibly gain efficiency without introducing bias. Also, in scenarios where the contamination lies not too far in the tails, the median estimator is also preferred, to ensure the “exclusion” of the outliers.

Moreover, comparing the unadjusted and adjusted estimators created a discussion about when observations can be classified as outliers, provided that it differs between the two types of estimators: small percentages of bad leverage points where the outcome value lies closely to its mean, would not get noted as an outlier and would hardly affect the difference-in-means estimate, while it can have a significant impact on the OLS estimate. Moreover, the different types of outliers could not even be classified without measuring/including the explanatory variables, whereas this can affect the applied economic/social/scientific inference greatly. Therefore, my personal preference would go to regression-adjusted approaches.

The most robust regression-adjusted estimator is the MM-estimator: It outperforms all other estimators (both in terms of bias and variance) and is very successful in removing the most strong and influential outliers. As a recommendation from this study, it would be to apply the MM-estimator in analysis on non-simulated data, where the presence of outliers is likely. However, one must not forget to check the required assumptions, such as the normality/symmetry of error terms. Therefore, a combination of various methods can be helpful when evaluating data. Moreover, the MM-estimator does not remove good leverage points from its analysis and it is therefore likely that it underestimates the treatment effect variance. Interpreting the results of just-significant estimates should therefore be done with caution and a critical eye could be laid on the standardised residuals-Mahalanobis distance plot (using the MM-estimator and the MCD estimator) to be mindful about the possible presence of good leverage points.

Additionally, the poor performance of the KW-estimator (despite being optimal robust) can be explained by the usage of the $DFFITS$-values for computing the weights: e.g. in the presence of only one bad leverage point, the $DFFITS$-value of this outlier is very large (and very small for all most other observations), significantly down-weighting this outlier. However, when there are multiple bad leverage points, $\hat{\beta}(i_{blp})$ might not be much different from $\hat{\beta}$, because the remaining bad leverage points still exert a strong influence on $\hat{\beta}$. As a result, the $DFFITS$-values for all bad leverage points individually are only small, therefore not down-weighting the outliers (enough).

Suggestions for further research would be to extend the simulation study, to dive deeper into the (difference-in-intercepts-)MM-estimator and to revisit the papers using the NSW data set. Firstly, it would be interesting to further investigate the differences in bias and variance between the difference-in-means estimator and OLS. Both IFs and CVFs are unbounded, but this gives little information about practical implications. For example, it is possible for the OLS variance to exceed that of the difference-in-means estimator, even when this is not the case in this thesis' simulation. It is interesting to further understand the implications of unbounded IFs and CVFs in different experimental settings (e.g. number of data points, initial sample distributions, level of contamination, location of point mass contamination) and to find possible relations between them. Secondly, Table 3 in Section 7 showed a discrepancy between the value and variance estimate of the regression-adjusted MM-estimators and the difference-in-intercepts variant, whereas this was not found for OLS. A better understanding of (bootstrapping) the MM-estimator and its intermediate steps can potentially explain this difference. More knowledge about the differences in estimates can further contribute to the area of analysing the effect of outliers on (robust) estimators. Lastly, it can be interesting to replicate the multiple studies using the NSW data (and the comparative data sets perhaps as well) using robust methods, and to evaluate if the practical inference still holds true.

9 References

Angrist, J. D. (2004). Treatment Effect Heterogeneity in Theory and Practice. *The Economic Journal*, 114(494):C52–C83.

Austin, P. C. and Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, 34(28):3661–3679.

Baharudin, Z. A., Ahmad, N. A., Fernando, M., Cooray, V., and Mäkelä, J. (2012). Comparative study on preliminary breakdown pulse trains observed in Johor, Malaysia and Florida, USA. *Atmospheric Research*, 117:111–121.

Barnett, V., Lewis, T., et al. (1994). *Outliers in statistical data*, volume 3. New York: John Wiley & Sons.

Berkhemer, O. A., Fransen, P. S., Beumer, D., Van Den Berg, L. A., Lingsma, H. F., Yoo, A. J., Schonewille, W. J., Vos, J. A., Nederkoorn, P. J., Wermer, M. J., et al. (2015). A randomized trial of intraarterial treatment for acute ischemic stroke. *The New England Journal of Medicine*, 372:11–20.

Blair, G., Cooper, J., Coppock, A., Humphreys, M., Sonnet, L., Fultz, N., Medina, L., and Lenth, R. (2022). *Fast Estimators for Design-Based Inference*. R Foundation for Statistical Computing.

Canavire-Bacarreza, G., Castro Peñarrieta, L., and Ugarte Ontiveros, D. (2021). Outliers in semi-parametric estimation of treatment effects. *Econometrics*, 9(2):19.

Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association*, 96:1022–1030.

Criscuolo, C., Martin, R., Overman, H., and Van Reenen, J. (2012). The causal effects of an industrial policy. Technical report, National Bureau of Economic Research.

Croux, C., Dhaene, G., and Hoorelbeke, D. (2004). Robust standard errors for robust estimators. *CES-Discussion Paper Series (DPS) 03.16*, pages 1–20.

Deaton, A. and Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210:2–21. Randomized Controlled Trials and Evidence-based Policy: A Multidisciplinary Dialogue.

Dechartres, A., Trinquart, L., Boutron, I., and Ravaud, P. (2013). Influence of trial sample size on treatment effect estimates: Meta-epidemiological study. *BMJ*, 346.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062.

Dehejia, R. H. and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1):151–161.

Diamond, A. and Sekhon, J. S. (2013). Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *The Review of Economics and Statistics*, 95(3):932–945.

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

Fisher, R., A. (1932). *Statistical Methods for Research Workers.* Oliver and Boyd.

Fisher, R. A. (1936). Design of experiments. *British Medical Journal*, 1(3923):554.

Flavin, M. (1999). Robust estimation of the joint consumption/asset demand decision. Available at `https://www.nber.org/papers/w7011`. Published by the National Bureau of Economic Research Cambridge, Massachusetts, USA.

Freedman, D. A. (2006). On the So-Called "Huber Sandwich Estimator" and "Robust Standard Errors". *The American Statistician*, 60(4):299–302.

Freedman, D. A. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193.

Grentzelos, C., Caroni, C., and Barranco-Chamorro, I. (2021). A comparative study of methods to handle outliers in multivariate data analysis. *Computational and Mathematical Methods*, 3(3):e1129.

Hampel, F. (1968). *Contributions to the Theory of Robust Estimation.* PhD thesis, University of California. Available at https://www.proquest.com/docview/302315485?pq-origsite=gscholarfromopenview=true.

Hampel, F. R. (1971). A General Qualitative Definition of Robustness. *The Annals of Mathematical Statistics*, 42(6):1887–1896.

Hampel, F. R. (1974). The Influence Curve and Its Role in Robust Estimation. *Journal of the American Statistical Association*, 69(346):383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions.* New York: John Wiley & Sons.

Hennig, C. (2004). Breakdown points for maximum likelihood estimators of location–scale mixtures. *The Annals of Statistics*, 32(4):1313 – 1340.

Hodge, V. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22:85–126.

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

Hollingsworth, J. M., Miller, D. C., Daignault, S., and Hollenbeck, B. K. (2006). Rising Incidence of Small Renal Masses: A Need to Reassess Treatment Effect. *JNCI: Journal of the National Cancer Institute*, 98(18):1331–1334.

Hsu, D. and Sabato, S. (2014). Heavy-tailed regression with a generalized median-of-means. In *International Conference on Machine Learning*, pages 37–45. PMLR.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1: Statistics, pages 221–233. University of California Press, Berkeley, CA.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust statistics.* New York: John Wiley & Sons.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 76(1):243–263.

Johansson, P. and Palme, M. (2002). Assessing the effect of public policy on worker absenteeism. *The Journal of Human Resources*, 37(2):381–409.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4).

Kempthorne, O. (1952). *The design and analysis of experiments.* New York: John Wiley & Sons.

King, R. D., Massoglia, M., and MacMillan, R. (2007). The context of marriage and crime: Gender, the propensity to marry, and offending in early adulthood. *Criminology*, 45(1):33–65.

Krasker, W. S. (1980). Estimation in linear regression models with disparate data points. *Econometrica*, 48(6):1333–1346.

Krasker, W. S. and Welsch, R. E. (1982). Efficient bounded-influence regression estimation. *Journal of the American Statistical Association*, 77(379):595–604.

Kurz, C. F. (2022). Augmented inverse probability weighting and the double robustness property. *Medical Decision Making*, (2):157–167.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620.

Lei, L. and Ding, P. (2020). Regression adjustment in completely randomized experiments with a diverging number of covariates. *Biometrika*, 108(4):815–828.

Li, J., Handorf, E., Bekelman, J., and Mitra, N. (2016). Propensity score and doubly robust methods for estimating the effect of treatment on censored cost. *Statistics in Medicine.*, 12(35):1985–1999.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: reexamining freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318.

Liu, H. and Yang, Y. (2020). Regression-adjusted average treatment effect estimates in stratified randomized experiments. *Biometrika*, 107(4):935–948.

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Rückstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E. L. T., and di Palma, M. A. (2023). *Basic Robust Statistics.* R Foundation for Statistical Computing.

Maronna, R., Bustos, O., and Yohai, V. (1979). *Smoothing Techniques for Curve Estimation: Bias- and efficiency-robustness of general M-estimators for regression with random carriers*, volume 757, pages 91–116.

Mehta, A. (2023). Robust estimator for linear regression. *SAS Institute.* Retrieved from https://support.sas.com/resources/papers/proceedings-archive/SUGI91/Sugi-91-253

Naci, H. and Ioannidis, J. P. (2015). Comparative effectiveness of exercise and drug interventions on mortality outcomes: Metaepidemiological study. *BMJ*, 49(21):1414–1422.

Negi, A. and Wooldridge, J. M. (2021). Revisiting regression adjustment in experiments with heterogeneous treatment effects. *Econometric Reviews*, 40(5):504–534.

Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. [Translated in Statistical Science (1990)]. *Annals of Agricultural Sciences*, 10:1–51.

R Core Team (2023). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rocke, D. M. and Woodruff, D. L. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*, 91(435):1047–1061.

Ronchetti, E. and Rousseeuw, P. J. (1985). Change-of-variance sensitivities in regression analysis. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 68(4):503–519.

Rousseeuw, P. (1981a). Infinitesimal criteria in robust estimation of location. *JORBEL-Belgian Journal of Operations Research, Statistics, and Computer Science*, 21(4):24–42.

Rousseeuw, P. and Wagner, J. (1994). Robust regression with a distributed intercept using least median of squares. *Computational Statistics & Data Analysis*, 17(1):65–76.

Rousseeuw, P. J. (1981b). A new infinitesimal approach to robust estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 56(1):127–132.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880.

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

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283.

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 Hubert, M. (2018). Anomaly detection by robust statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(2):e1236.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust regression and outlier detection.* New York: John Wiley & Sons.

Rousseeuw, P. J. and van Driessen, K. (2005). Computing LTS Regression for Large Data Sets. *Data Mining and Knowledge Discovery*, 12:29–45.

Rousseeuw, R. and Yohai, V. (1984). Robust Regression by Means of S-Estimators. *Robust and Nonlinear Time Series Analysis*, 26:256–272.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2(1):1–26.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, pages 34–58.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Salibian-Barrera, M. and Yohai, V. J. (2006). A fast algorithm for S-regression estimates. *Journal of Computational and Graphical Statistics*, 15(2):414–427.

Sampson, R. J., Laub, J. H., and Wimer, C. (2006). Does marriage reduce crime? A counterfactual approach to within-individual causal effects. *Criminology*, 44(3):465–508.

Shambaugh, J. C. (2004). The Effect of Fixed Exchange Rates on Monetary Policy. *The Quarterly Journal of Economics*, 119(1):301–352.

Siegel, A. F. (1982). Robust regression using repeated medians. *Biometrika*, 69(1):242–244.

Smith, J. A. and Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(1):305–353. Experimental and non-experimental evaluation of economic policy and models.

Stolberg, C. R., Mundbjerg, L. H., Funch-Jensen, P., Gram, B., Bladbjerg, E. M., and Juhl, C. B. (2018). Effects of gastric bypass surgery followed by supervised physical training on inflammation and endothelial function. *Atherosclerosis*, 273:37–44.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, 33(1):1–67.

Tukey, J. W. (1977). *Exploratory data analysis*, volume 2. Addison-Wesley Publishing Company.

Verardi, V. and Croux, C. (2009). Robust regression in stata. *The Stata Journal*, 9(3):439–453.

White, H. (1980a). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838.

White, H. (1980b). Using least squares to approximate unknown regression functions. *International Economic Review*, 21(1):149–170.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 15(2).

Yusuf, S., Bosch, J., Dagenais, G., Zhu, J., Xavier, D., Liu, L., Pais, P., Lopez-Jaramillo, P., Leiter, L. A., Dans, A., and Avezum, A. (2016). Cholesterol lowering in intermediate-risk persons without cardiovascular disease. *The New England Journal of Medicine*, 374(21).

Zaman, A., Rousseeuw, P. J., and Orhan, M. (2001). Econometric applications of high-breakdown robust regression techniques. *Economics Letters*, 71(1):1–8.

Zeileis, A., Lumley, T., Graham, N., and Koell, S. (2022). *Robust Covariance Matrix Estimators.* R Foundation for Statistical Computing.

Zhelonkin, M. (2013). *Robustness in Sample Selection Models.* PhD thesis, Université de Genève, 1205 Genève, Switserland. Available at `https://archive-ouverte.unige.ch/unige:27996`.

A IF & CVF derivations

A.1 Derivation IF one-step M-estimators

The derivation starts with the system of equations $\Psi_1\{z, S(F)\}dF = 0$ and its contaminated version $\Psi_1\{z, S(F_\varepsilon)\}dF_\varepsilon = 0$. Taking the derivative with respect to ε at $\varepsilon = 0$ gives

$$\begin{aligned}
&\frac{\partial}{\partial\varepsilon}\int \Psi_1\{z, S(F_\varepsilon)\}dF_\varepsilon\Big|_{\varepsilon=0} = 0\\
&\Leftrightarrow \frac{\partial}{\partial\varepsilon}\Big[(1-\varepsilon)\int \Psi_1\{z, S(F_\varepsilon)\}dF + \varepsilon\int \Psi_1\{z, S(F_\varepsilon)\}d\Delta_z\Big]\Big|_{\varepsilon=0} = 0\\
&\Leftrightarrow \frac{\partial}{\partial\varepsilon}\int \Psi_1\{z, S(F_\varepsilon)\}dF\Big|_{\varepsilon=0} - \int \Psi_1\{z, S(F)\}dF + \Psi_1\{z, S(F)\} = 0\\
&\Leftrightarrow \int \frac{\partial}{\partial\theta}\Psi_1(z,\theta)dF IF(z; S, F) + \Psi_1\{z, S(F)\} = 0\\
&\Leftrightarrow IF(z; S, F) = \Big(-\int \frac{\partial}{\partial\theta}\Psi_1(z,\theta)dF\Big)^{-1}\Psi_1\{z, S(F)\},
\end{aligned}$$

where the derivative with respect to θ is evaluated at $\theta = S(F)$.

A.2 Derivation IF mean

The derivation starts with the notion that the mean score function $\Psi\{z, S(F)\} \Rightarrow \int(y - \mu)dF = 0 \Leftrightarrow \text{argmin}_\mu \sum_{i=1}^{N}(y_i - \mu)^2$, using a squared loss function. Then, the derivation start with the derivative of the contaminated score function:

$$\begin{aligned}
&\frac{\partial}{\partial\varepsilon}\int \Psi\{z, S(F_\varepsilon)\}dF_\varepsilon\Big|_{\varepsilon=0} = \frac{\partial}{\partial\varepsilon}\int \{y - S(F_\varepsilon)\}dF_\varepsilon\Big|_{\varepsilon=0} = 0\\
&\Leftrightarrow \frac{\partial}{\partial\varepsilon}\Big[\int (1-\varepsilon)\{y - S(F_\varepsilon)\}dF + \varepsilon\int \{y - S(F_\varepsilon)\}d\Delta_z\Big]\Big|_{\varepsilon=0} = 0\\
&\Leftrightarrow -\int \{y - S(F_\varepsilon)\}dF\Big|_{\varepsilon=0} + \frac{\partial}{\partial\varepsilon}\int \{y - S(F_\varepsilon)\}dF\Big|_{\varepsilon=0} + y - S(F) = 0\\
&\Leftrightarrow -\int \{y - S(F)\}dF + \frac{\partial}{\partial\varepsilon}\Big[\int ydF - \int S(F_\varepsilon)dF\Big]\Big|_{\varepsilon=0} + y - \mu = 0\\
&\Leftrightarrow -\frac{\partial}{\partial\varepsilon}S(F_\varepsilon)\int dF\Big|_{\varepsilon=0} + y - \mu = 0\\
&\Leftrightarrow IF(z; S, F) = y - \mu
\end{aligned}$$

A.3 Derivation CVF difference-in-means estimator

The derivation starts by using the two-stage M-estimator CVF expression in (3.7) by Zhelonkin (2013), described in Theory Section 3.5.3. This derivation starts by analysing all terms individually and then combining them to get the final CVF expression for the difference-in-mean estimator. To start off: $a(z) = \Psi_2 = \mu_1 - \mu_0 - \tau$ and $b(z) = \begin{bmatrix}1 & -1\end{bmatrix}\Psi_1\{z^{(1)}, S(F)\}$.
Matrix A equals

$$
\begin{aligned}
A =& \frac{\partial}{\partial h}\Psi_2\big\{z^{(2)}, h, T(F)\big\}\frac{\partial h(z^{(1)}, s)}{\partial s} IF(z; S, F) \\
&+ \frac{\partial}{\partial \theta}\Psi_2\Big[z^{(2)}, h\big\{z^{(1)}, S(F)\big\}, \theta\Big] IF(z; T, F) \\
&= \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} y_1 - \mu_1 \\ y_0 - \mu_0 \end{bmatrix} - (y_1 - y_0 - \tau) \\
&= -(\mu_1 - \mu_0 - \tau) \\
&= -\Psi_2 = -a(z).
\end{aligned}
$$

For matrix B, elements R_2, R_1 and $D^{(1)}$ are required, which will be derived first:

matrix $D^{(1)}$ has elements

$$
\begin{aligned}
D^{(1)}_{ij} =& \Big(\frac{\partial}{\partial \theta}\frac{\partial \Psi_{1i}(z^{(1)}, \theta)}{\partial \theta_j}\Big)^T IF(z; S, F) = 0 \\
& \text{since } \frac{\partial \Psi_1}{\partial \theta} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \text{ and therefore } \frac{\partial}{\partial \theta}\frac{\partial \Psi_1}{\partial \theta} = 0,
\end{aligned}
$$

making it a full zero-matrix; matrix R_1 has elements

$$
\begin{aligned}
R^{(1)}_{ij} =& \Big(\frac{\partial}{\partial h}\frac{\partial \Psi_{2i}\big\{z^{(2)}, h, T(F)\big\}}{\partial h_j}\Big)^T \frac{\partial h(z^{(1)}, s)}{\partial s} IF(z; S, F) \\
&+ \frac{\partial}{\partial \theta}\frac{\partial \Psi_{2i}\big\{z^{(2)}, h, \theta\big\}}{\partial h_j} IF(z; T, F) = 0 \\
& \text{since } \frac{\partial \Psi_2}{\partial h} = 1, \text{ and therefore } \frac{\partial}{\partial h}\frac{\partial \Psi_2}{\partial h} = \frac{\partial}{\partial \theta}\frac{\partial \Psi_2}{\partial h} = 0,
\end{aligned}
$$

making it a full zero-matrix, as well; and lastly matrix R_2 has elements:

$$
\begin{aligned}
R^{(2)}_{ij} =& \Big(\frac{\partial}{\partial s}\frac{\partial h_i(z^{(1)}, s)}{\partial s_j}\Big)^T IF(z; S, F) = 0 \\
& \text{since } \frac{\partial h}{\partial s} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}^T, \text{ and therefore } \frac{\partial}{\partial s}\frac{\partial h}{\partial s} = 0,
\end{aligned}
$$

also making this last sub-matrix a zero-matrix. Now, matrix B can be constructed: Note that since the three matrices R_2, R_1 and $D^{(1)}$ are all zero-matrices, terms including those matrices can immediately be omitted

$$
\begin{aligned}
B =& \int R_1 \frac{\partial}{\partial s} h(z^{(1)}, s) dF IF(z; S, F) + \int \frac{\partial}{\partial h}\Psi_2\big\{z^{(2)}, h, T(F)\big\} R_2 dF IF(z; S, F) \\
&- \int \frac{\partial}{\partial h}\Psi_2(z^{(2)}, h, T(F)) \frac{\partial}{\partial s} h(z^{(1)}, s) dF M_1^{-1} \int D^{(1)} dF IF(z; S, F) \\
&+ \frac{\partial}{\partial h}\Psi_2\big\{z^{(2)}, h, T(F)\big\}\frac{\partial}{\partial s} h(z^{(1)}, s) IF(z; S, F) \\
=& \frac{\partial}{\partial h}\Psi_2\big\{z^{(2)}, h, T(F)\big\}\frac{\partial}{\partial s} h(z^{(1)}, s) IF(z; S, F) \\
& \begin{bmatrix} 1 \\ -1 \end{bmatrix}^T IF(z; S, F) = b(z).
\end{aligned}
$$

Now, only $D^{(2S)}$ should be computed to derive the CVF: matrix $D^{(2S)}$ has elements:

$$\begin{aligned}D^{(2S)}_{ij} &= \Big(\frac{\partial}{\partial h}\frac{\partial\Psi_{2i}\big\{z^{(2)},h,\theta\big\}}{\partial\theta_j}\Big)^T\frac{\partial h(z^{(1)},s)}{\partial s}IF(z;S,F)\\ &+\Big(\frac{\partial}{\partial\theta}\frac{\partial\Psi_{2i}\big\{z^{(2)},h,\theta\big\}}{\partial\theta_j}\Big)^T IF(z;T,F)\\ &\text{since } \frac{\partial\Psi_2}{\partial\theta}=-1, \text{ and therefore } \frac{\partial}{\partial h}\frac{\partial\Psi_2}{\partial\theta}=\frac{\partial}{\partial\theta}\frac{\partial\Psi_2}{\partial\theta}=0,\end{aligned}$$

also making it a zero-matrix. Now, the generic two-stage M-estimator CVF can be shrunk down through omitting the zero-matrices terms and combining lines 3 and 4, since both $a(z)$ and $b(z)$ are scalars:

$$\begin{aligned}CVF(z;S,T,F) &= V(T,F)-M^{-1}\Big(\int D^{(2S)}dF\Big)V(T,F)\\ &-M^{-1}\Big(\frac{\partial}{\partial\theta}\Psi_2\Big[z^{(2)},h\big\{z^{(1)},S(F)\big\},\theta\Big]\Big)V(T,F)\\ &+M^{-1}\Big(\int\big\{Aa(z)^T+Ba(z)^T+Ab(z)+Bb(z)^T\big\}dF\Big)M^{-1}\\ &+M^{-1}\Big(\int\big\{a(z)A^T+b(z)A^T+a(z)B^T+b(z)B^T\big\}dF\Big)M^{-1}\\ &+M^{-1}\Big(a(z)a(z)^T+a(z)b(z)^T+b(z)a(z)^T+b(z)b(z)^T\Big)M^{-1}\\ &-V(T,F)\Big(\int D^{(2S)}dF+\frac{\partial}{\partial\theta}\Psi_2\Big[z^{(2)},h\big\{z^{(1)},S(F)\big\},\theta\Big]\Big)M^{-1}\\ &=V(T,F)-0+V(T,F)\\ &+2\int\big\{-a(z)A+b(z)^2\big\}dF+\big\{a(z)+b(z)\big\}^2+V(T,F)\\ &=3V(T,F)+2\int\big\{-a(z)A+b(z)^2\big\}dF+\big\{a(z)+b(z)\big\}^2\end{aligned}$$

where each term in the second (final) step represents a full line in the general formula (first step).

A.4 Derivation CVF one-stage regression-adjusted estimator

Generally, the derivation and final outcome are very similar to these of Appendix A.3 for the derivation of the difference-in-means CVF. The outcome therefore is also similar, only substituting for different score- and IF-functions. Three notes to add to the previous derivation are:

1. Matrix A still equals $-\Psi_2$, but it deserved a more general derivation:

$$\begin{aligned}A &= \frac{\partial}{\partial h}\Psi_2\big\{z^{(2)},h,T(F)\big\}\frac{\partial h(z^{(1)},s)}{\partial s}IF(z;S,F)\\ &+\frac{\partial}{\partial\theta}\Psi_2\Big[z^{(2)},h\big\{z^{(1)},S(F)\big\},\theta\Big]IF(z;T,F)\\ &=\begin{bmatrix}0 & 1 & 0^T & 0^T\end{bmatrix}IF(z;S,F)-(\Psi_2+\begin{bmatrix}0 & 1 & 0^T & 0^T\end{bmatrix}IF(z;S,F))\\ &=-\Psi_2=-a(z).\end{aligned}$$

2. $\frac{\partial \Psi_1}{\partial \theta}$ as seen in the steps for computing $D^{(1)}$ now equals $-zz^T$ with $z = \begin{bmatrix} 1 \\ T \\ x \\ T(x-\mu_x) \end{bmatrix}$.

   However, then still $\frac{\partial}{\partial \theta}\frac{\partial \Psi_1}{\partial \theta} = 0$, so the $D^{(1)}$ remains a zero-matrix.

3. $\frac{\partial h}{\partial s}$ as seen in the steps for computing $R_{ij}^{(2)}$ is $\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}^T$, but $\frac{\partial}{\partial s}\frac{\partial h}{\partial s} = 0$ remains. In similar fashion, $B = b(z)$, still.

A.5 Derivation CVF difference-in-intercepts estimator

Generally, the derivation and final outcome are very similar to these of Appendix A.3 for the derivation of the difference-in-means CVF. Just like the regression adjusted version in Appendix A.4, the outcome therefore is also similar, only substituting for different score- and IF-functions. Two similar notes to add to the difference-in-means CVF derivation are:

1. $\frac{\partial \Psi_1}{\partial \theta}$ as seen in the steps for computing $D^{(1)}$ now equals $\begin{bmatrix} -\begin{bmatrix} 1 \\ x_1 - \mu_{x1} \end{bmatrix} \begin{bmatrix} 1 & x_1 - \mu_{x1} \end{bmatrix} \\ -\begin{bmatrix} 1 \\ x_0 - \mu_{x0} \end{bmatrix} \begin{bmatrix} 1 & x_0 - \mu_{x0} \end{bmatrix} \end{bmatrix}$.

   However, then still $\frac{\partial}{\partial \theta}\frac{\partial \Psi_1}{\partial \theta} = 0$, so the $D^{(1)}$ remains a zero-matrix.

2. $\frac{\partial h}{\partial s}$ as seen in the steps for computing $R_{ij}^{(2)}$ is $\begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix}^T$, but $\frac{\partial}{\partial s}\frac{\partial h}{\partial s} = 0$ remains. In similar fashion, $B = b(z)$, still.

B Additional tables and figures simulation study

B.1 Estimation results for the full set of covariates

Table 4: Results regression-adjusted estimators

	OLS					**MM**					**KW**				
	int.	**T**	**x1**	**x2**	**x3**	**int.**	**T**	**x1**	**x2**	**x3**	**int.**	**T**	**x1**	**x2**	**x3**
Clean	-0.002	0.253	1.000	1.001	1.000	-0.001	0.252	1.001	1.000	1.000	-0.000	0.250	1.001	0.998	1.000
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.000)
BLP2 (T&C)	-0.101	0.253	0.919	0.852	0.830	-0.001	0.252	1.001	1.000	1.000	-0.058	0.253	0.954	0.917	0.903
	(0.002)	(0.004)	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)	(0.003)	(0.001)	(0.001)	(0.001)
BLP2 (T)	-0.001	0.150	0.964	0.941	0.931	0.001	0.247	1.002	0.999	1.001	-0.001	0.200	0.980	0.972	0.965
	(0.001)	(0.004)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)	(0.001)	(0.001)	(0.001)
BLP2 (C)	-0.106	0.355	0.948	0.906	0.895	-0.002	0.256	0.999	1.001	1.002	-0.058	0.307	0.972	0.950	0.945
	(0.002)	(0.003)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)	(0.001)	(0.001)	(0.001)
BLP1 (T&C)	-0.057	0.252	0.954	0.918	0.904	-0.001	0.253	1.001	1.000	1.000	-0.041	0.253	0.968	0.942	0.932
	(0.001)	(0.003)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.007)
BLP1 (T)	-0.001	0.194	0.979	0.968	0.961	0.001	0.247	1.002	0.999	1.001	-0.001	0.214	0.985	0.980	0.975
	(0.001)	(0.003)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)	(0.001)	(0.001)	(0.001)
BLP1 (C)	-0.061	0.309	0.970	0.947	0.942	-0.002	0.256	0.999	1.001	1.002	-0.042	0.291	0.979	0.964	0.961
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
GLP (T&C)	-0.002	0.252	1.000	1.001	1.000	-0.001	0.253	1.001	1.000	1.000	-0.002	0.252	1.000	1.003	1.000
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
GLP (T)	-0.001	0.252	0.997	1.003	1.000	0.001	1.001	0.247	0.999	1.001	-0.001	0.252	0.997	1.003	1.000
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
GLP (C)	-0.002	0.251	0.999	1.000	1.002	-0.003	0.256	0.999	1.001	1.001	-0.002	0.251	0.999	1.000	1.002
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
VERT (T&C)	0.045	0.252	1.000	1.001	1.000	-0.001	0.253	1.001	1.000	1.000	0.045	0.252	1.000	1.001	1.000
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
VERT (T)	-0.001	0.300	0.997	1.003	1.000	0.001	0.248	1.002	0.999	1.001	-0.001	0.298	0.997	1.003	1.000
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
VERT (C)	0.046	0.202	0.999	1.000	1.002	-0.002	0.255	0.999	1.001	1.002	0.046	0.202	0.999	1.000	1.002
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)

All results are significant for $\alpha = 0.01$. Items in parentheses are estimator variances

Table 5: Results individual regressions difference-in-intercepts OLS-estimator

	OLS (T)				**OLS (C)**			
	int.	**x1**	**x2**	**x3**	**int.**	**x1**	**x2**	**x3**
Clean	0.250	0.999	1.000	1.002	-0.002	1.001	1.001	0.998
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
BLP2 (T&C)	0.153	0.918	0.853	0.834	-0.100	0.920	0.852	0.827
	(0.002)	(0.003)	(0.004)	(0.004)	(0.001)	(0.002)	(0.003)	(0.003)
BLP2 (T)	0.153	0.917	0.856	0.834	-0.001	0.998	1.002	1.000
	(0.003)	(0.003)	(0.005)	(0.005)	(0.001)	(0.001)	(0.001)	(0.001)
BLP2 (C)	0.249	0.999	0.999	1.004	-0.103	0.916	0.849	0.827'
	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.003)	(0.003)
BLP1 (T&C)	0.196	0.953	0.918	0.908	-0.057	0.955	0.918	0.902
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)
BLP1 (T)	0.196	0.952	0.921	0.907	-0.001	0.998	1.002	1.000
	(0.002)	(0.002)	(0.003)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
BLP1 (C)	0.249	0.999	0.999	1.004	-0.059	0.952	0.916	0.904
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)
GLP (T&C)	0.250	0.999	1.000	1.002	-0.002	1.001	1.002	0.998
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
GLP (T)	0.250	0.997	1.003	1.000	-0.001	0.998	1.002	1.000
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
GLP (C)	0.249	0.999	0.999	1.004	-0.002	0.999	1.001	1.001
	(0.001)	(0.002)	(0.002)	(0.01)	(0.001)	(0.001)	(0.001)	(0.001)
VERT (T&C)	0.297	0.999	1.000	1.002	0.045	1.001	1.002	0.998
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
VERT (T)	0.297	0.997	1.003	1.000	-0.001	0.998	1.002	1.000
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
VERT (C)	0.249	0.999	0.999	1.004	0.046	0.999	1.001	1.001
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)

All results are significant for $\alpha = 0.01$. Items in parentheses are estimator variances

Table 6: Results individual regressions difference-in-intercepts MM-estimator

	MM (T)				**MM (C)**			
	int.	**x1**	**x2**	**x3**	**int.**	**x1**	**x2**	**x3**
Clean	0.251	1.001	0.999	1.001	-0.001	1.000	1.001	0.999
	(0.01)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
BLP2 (T&C)	0.2517	1.001	0.999	1.001	-0.001	1.000	1.001	0.999
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
BLP2 (T)	0.249	1.003	0.999	1.001	0.001	1.000	0.999	1.001
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
BLP2 (C)	0.253	0.997	1.001	1.002	-0.002	1.000	1.002	1.001
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
BLP1 (T&C)	0.252	1.001	0.999	1.001	-0.001	1.000	1.001	0.999
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
BLP1 (T)	0.249	1.003	0.999	1.001	0.001	1.000	0.999	1.001
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
BLP1 (C)	0.253	0.997	1.001	1.002	-0.002	1.000	1.002	1.001
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
GLP (T&C)	0.252	1.001	0.999	1.001	-0.001	1.000	1.001	0.999
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
GLP (T)	0.249	1.003	0.999	1.001	0.001	1.000	0.999	1.001
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
GLP (C)	0.253	0.997	1.001	1.002	-0.003	1.000	1.002	1.001
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
VERT (T&C)	0.252	1.001	0.999	1.001	-0.001	1.000	1.001	0.999
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
VERT (T)	0.249	1.003	0.999	1.001	0.001	1.000	0.999	1.001
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
VERT (C)	0.253	0.997	1.001	1.002	-0.002	1.000	1.002	1.001
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)

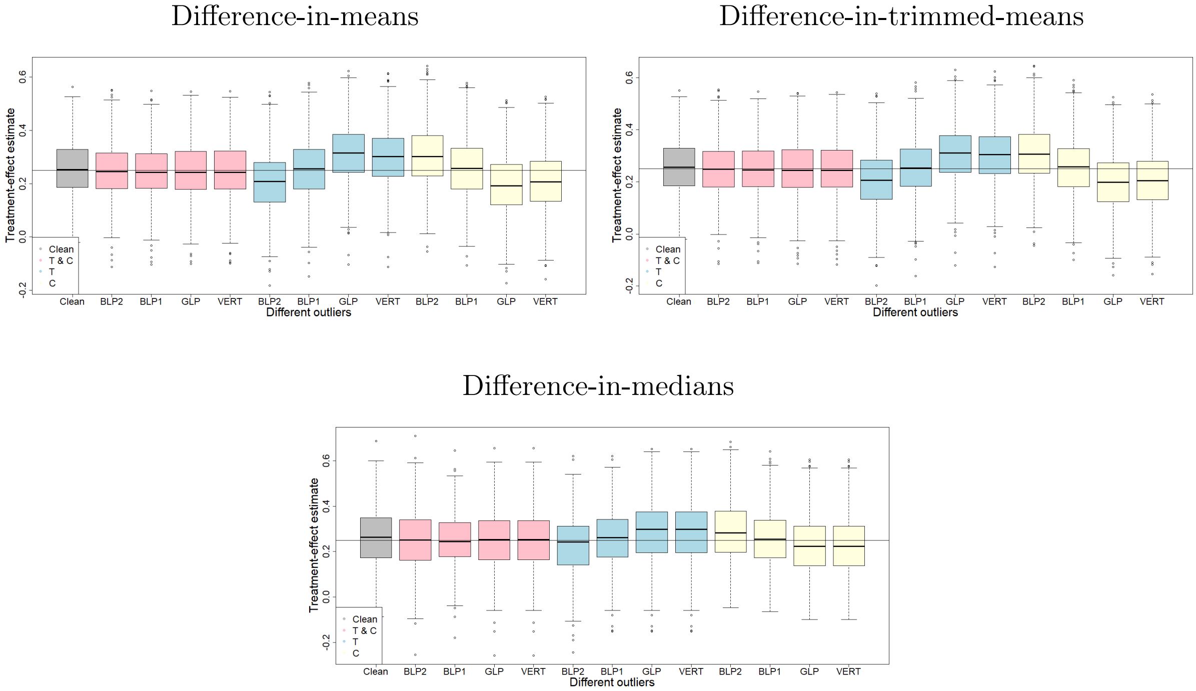
All results are significant for $\alpha = 0.01$. Items in parentheses are estimator variances

Table 7: Results individual regressions difference-in-intercepts KW-estimator

	KW (T)				**KW (C)**			
	int.	**x1**	**x2**	**x3**	**int.**	**x1**	**x2**	**x3**
Clean	0.250	1.001	0.999	1.000	-0.000	1.000	0.998	1.001
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
BLP2 (T&C)	0.200	0.957	0.924	0.915	-0.053	0.959	0.924	0.909
	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)
BLP2 (T)	0.200	0.955	0.927	0.914	-0.001	0.998	1.002	1.000
	(0.002)	(0.002)	(0.003)	(0.003)	(0.001)	(0.001)	(0.001)	(0.001)
BLP2 (C)	0.249	0.999	0.999	1.004	-0.054	0.955	0.922	0.911
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
BLP1 (T&C)	0.215	0.969	0.946	0.941	-0.039	0.971	0.947	0.935
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
BLP1 (T)	0.215	0.967	0.949	0.940	-0.001	0.998	1.002	1.000
	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
BLP1 (C)	0.249	0.999	0.999	1.004	-0.039	0.969	0.945	0.937
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
GLP (T&C)	0.250	0.999	1.000	1.002	-0.002	1.001	1.002	0.998
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
GLP (T)	0.250	0.997	1.003	1.001	-0.001	0.998	1.002	1.000
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
GLP (C)	0.249	0.999	0.999	1.004	-0.002	0.999	1.001	1.001
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
VERT (T&C)	0.297	0.999	1.000	1.002	0.045	1.001	1.002	0.998
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
VERT (T)	0.297	0.997	1.003	1.001	-0.001	0.998	1.002	1.003
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
VERT (C)	0.249	0.999	0.999	1.004	0.047	0.999	1.001	1.001
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)

All results are significant for $\alpha = 0.01$. Items in parentheses are estimator variances

B.2 Alternative representation boxplots

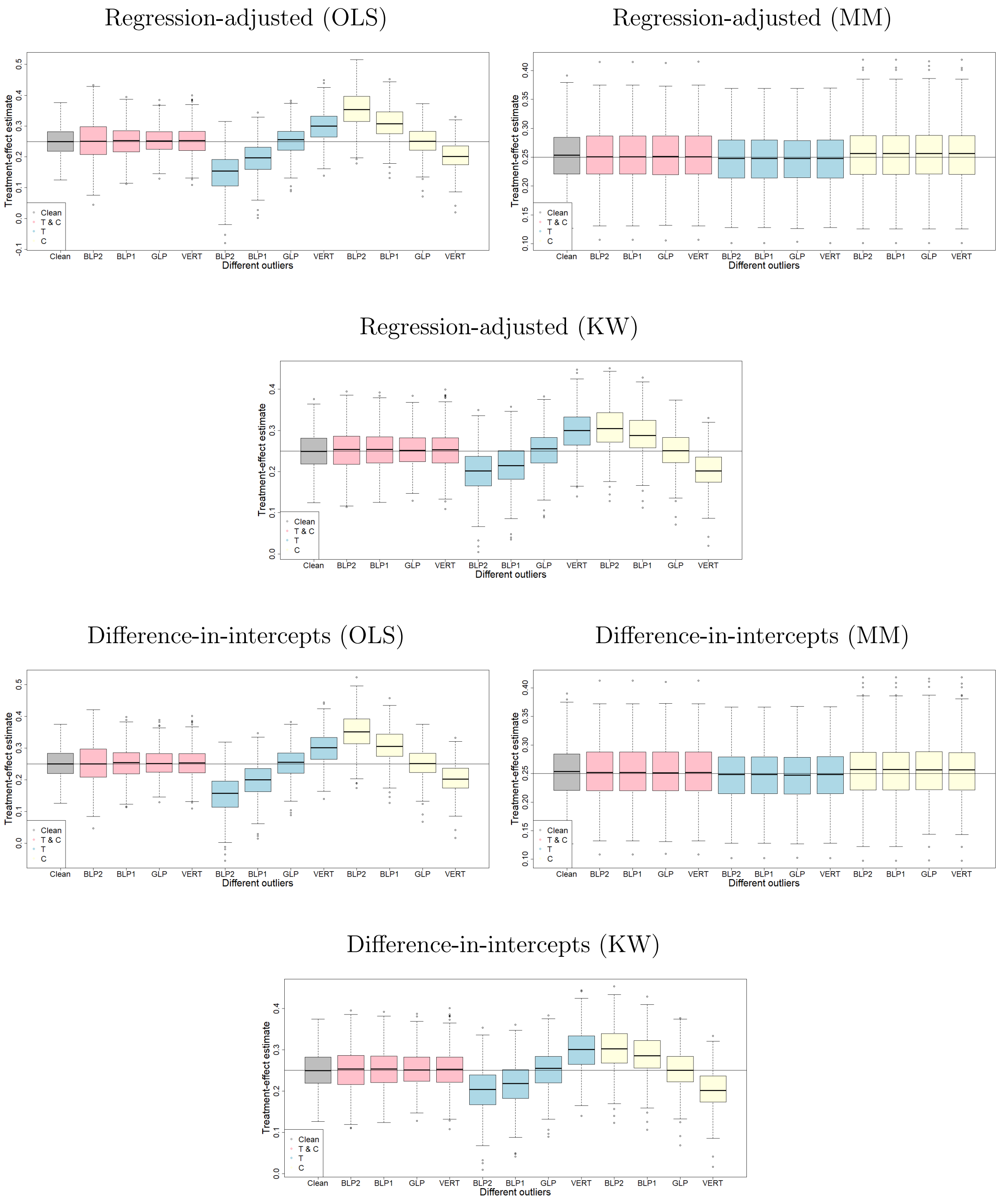


Difference-in-means

Difference-in-trimmed-means

Difference-in-medians

Figure 8: Boxplots difference-in-means based treatment effect estimators.



Regression-adjusted (OLS)

Regression-adjusted (MM)

Regression-adjusted (KW)

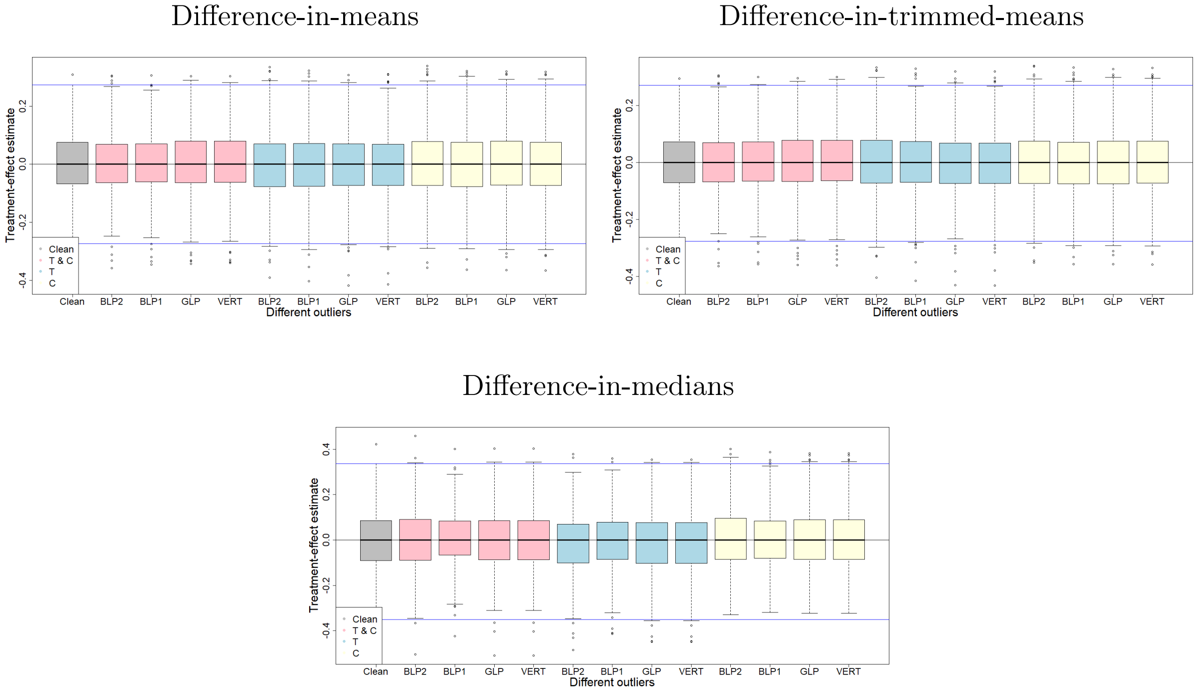
Difference-in-intercepts (OLS)

Difference-in-intercepts (MM)

Difference-in-intercepts (KW)

Figure 9: Boxplots treatment effect estimates including covariates.

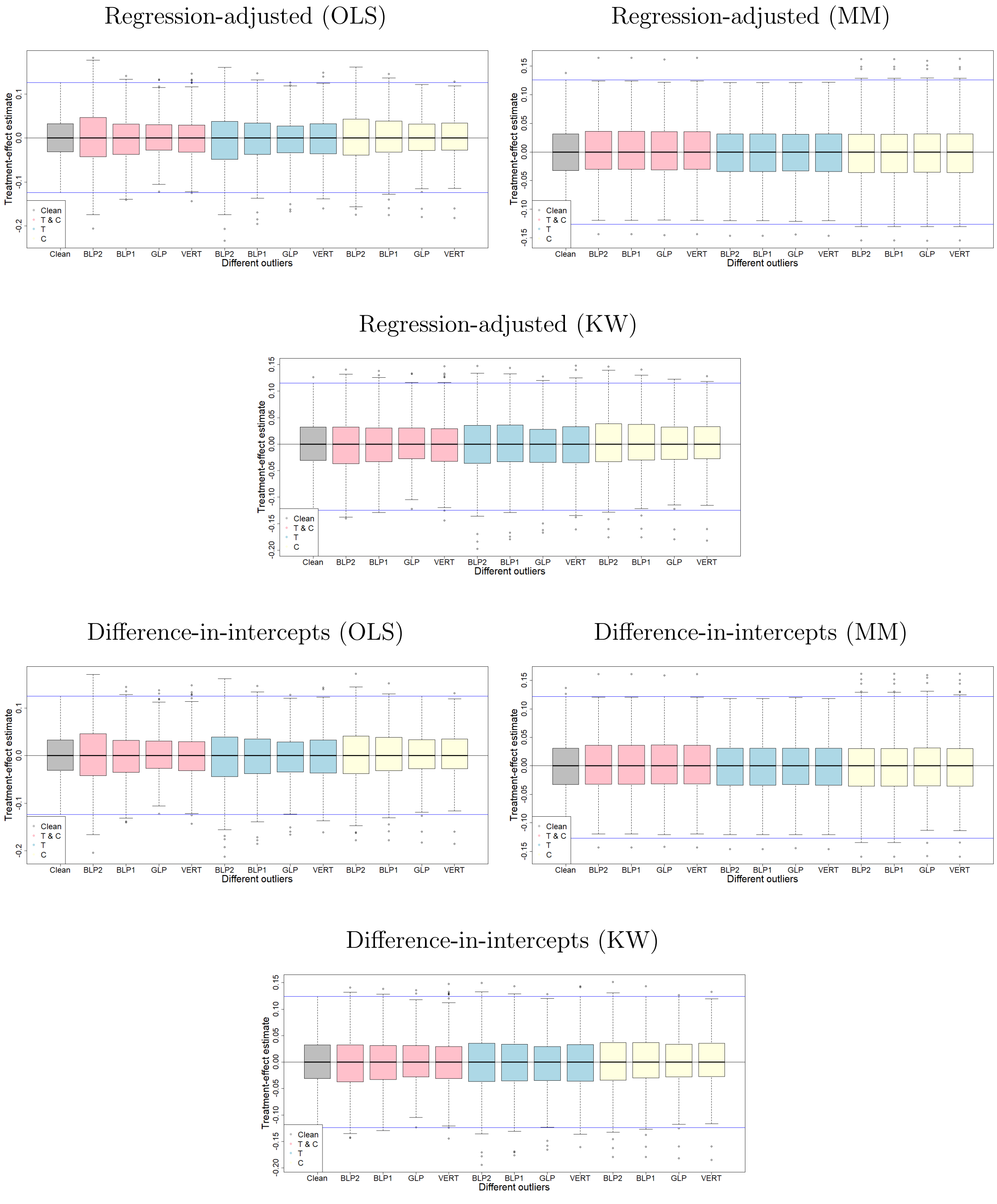
B.3 Centered boxplots


Difference-in-means

Difference-in-trimmed-means

Difference-in-medians

Figure 10: Centered boxplots difference-in-means based treatment effect estimators.



Regression-adjusted (OLS)

Regression-adjusted (MM)

Regression-adjusted (KW)

Difference-in-intercepts (OLS)

Difference-in-intercepts (MM)

Difference-in-intercepts (KW)

Figure 11: Centered boxplots treatment effect estimates including covariates.

C Additional tables and figures NSW data study

C.1 Estimation results for the full set of covariates

Table 8: Results regression-adjusted estimators

	Linear Regression			**Lin (2013)^3**		
	OLS	MM	KW^1	OLS	MM	KW^1
constant	4,938	12,034***	5,100	5,125***	11,133***	8.8856**
	(3,699)	(2,761)	(3,593)	(275)	(3,589)	(3915)
T	791	433	763	785	413	755
	(456)	(401)	(475)	(484)	(404)	(472)
age	-34	-370**	-43	-101	-318	-102
	(197)	(167)	(194)	(211)	(199)	(211)
age^2	0.92	5.69**	1.09	1.77	5.44*	1.79
	(3.21)	(2.77)	(3.16)	(3.36)	(3.18)	(3.36)
education	213	-16	206	19	21	22
	(165)	(138)	(161)	(218)	(177)	(217)
black	-1,767**	-2,404***	$-1,731^{**}$	-2,319**	-2,558**	$-2,297^{**}$
	(768)	(767)	(740)	(1,081)	(1,047)	(1,070)
hispanic	-146	-834	-102	-306	-259	-282
	(985)	(934)	(960)	(1,320)	(1,265)	(1,309)
married	562	869	478	-23	173	-35
	(685)	(566)	(651)	(795)	(725)	(790)
nodegree	-532	-828	-534	-936	-1,107	-935
	(759)	(650)	(745)	(894)	(880)	(894)
T:age				204	-93	176
				(452)	(372)	(439)
T:age^2				-2.79	0.56	-2.36
				(7.45)	(6.27)	(7.28)
T:education				476	-38	452
				(348)	(292)	(341)
T:black				1,330	165	1,275
				(1,541)	(1,525)	(1,529)
T:hispanic				75	-1,646	40
				(2,028)	(1,872)	(2,021)
T:married				1,466	1,714	1,505
				(1,443)	(1,157)	(1,417)
T:nodegree				1,042	759	951
				(1,552)	(1,316)	(1,512)

1: The KW-estimators did not converge, but bounced between two local minima. Estimates of one minima are still included because the estimates were relatively close to each other and to show the similarity to the OLS estimates.

3: The covariates in the interaction term are demeaned

Treatment-effect estimates, robust standard errors in parentheses.*** p<0.01, ** p<0.05, * p<0.1

Table 9: Results individual regressions difference-in-intercepts estimators

	Difference-in-intercepts[3]					
	OLS (T)	OLS (C)	MM (T)	MM (C)	KW[1] (T)	KW[1] (C)
constant	5,909*** (404)	5,125*** (277)	4,654*** (444)	4,079*** (380)	5,877*** (390)	5,108*** (272)
age	103 (399)	-101 (246)	-411 (314)	-339* (202)	81 (392)	-106 (210)
age^2	-1.01 (6.71)	1.77 (4.11)	6.07 (5.41)	5.78* (3.24)	-0.65 (6.54)	1.87 (3.34)
education	495 (307)	19 (226)	12 (245)	25 (176)	477* (265)	33 (211)
black	-990 (1,332)	-2,319** (997)	-2,321** (1,137)	-2,620** (1,051)	-1,029 (1,090)	-2,171** (1,003)
hispanic	-231 (1,842)	-306 (1,266)	-1,864 (1,375)	-328 (1,257)	-341 (1,495)	-145 (1,246)
married	1,444 (1,123)	-22.65 (784)	1,846** (893)	203 (731)	1,367 (1,145)	-107 (763)
nodegree	107 (1,240)	-936 (936)	-345 (978)	-1,112 (886)	102 (1,249)	-923 (892)

1: The KW-estimators did not converge, but bounced between two local minima. Estimates of one minima are still included because the estimates were relatively close to each other and to show the similarity to the OLS estimates.
3: All covariates are demeaned
Treatment-effect estimates, robust standard errors in parentheses.*** p<0.01, ** p<0.05, * p<0.1

C.2 Alternative Mahalanobis distance plots

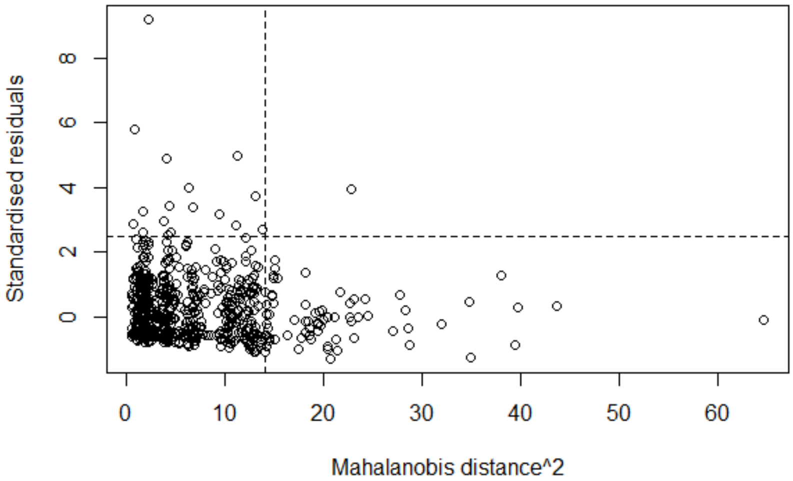


Figure 12: Standardised residuals plotted against the Mahalanobis distance squared, excluding the *agesquared* variable.

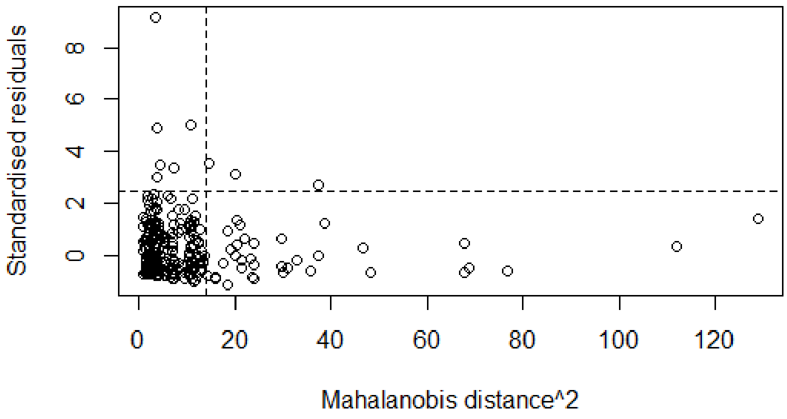


Figure 13: Standardised residuals plotted against the Mahalanobis distance squared, for the treatment group only.

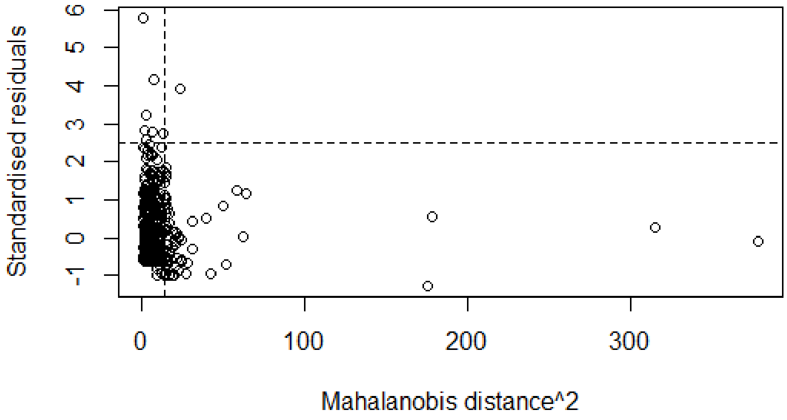


Figure 14: Standardised residuals plotted against the Mahalanobis distance squared, for the control group only.

C.3 Covariate information observations with largest standardised residuals

Table 10: Covariates 15 observations with the largest standardised residuals (bad leverage points & vertical outliers)

id	T	Age	Age^2	Educ.	Bl ack	Hisp.	Mar ried	No deg.	Inc. 1978	Out lier	St. res.	Mah. dist.^{2}	w
82	1	28	784	11	1	0	0	1	60,308	VERT	9.19	3.37	0
386	0	21	441	10	1	0	0	1	39,484	VERT	5.81	0.89	0
291	1	25	625	14	1	0	1	0	36,647	VERT	4.98	11.30	0
127	1	27	729	13	1	0	0	0	34,099	VERT	4.90	4.56	0
475	0	21	441	14	1	0	0	0	29,408	VERT	4.00	6.69	0
609	0	26	676	8	0	0	1	1	30,248	BLP	3.94	23.03	0
116	1	31	961	9	0	1	0	1	26,818	VERT	3.73	13.73	0
3	1	30	900	12	1	0	0	0	24,909	VERT	3.45	4.82	0
33	1	26	676	11	1	0	1	1	26,372	VERT	3.39	7.03	0
425	0	23	529	11	1	0	0	1	23,483	VERT	3.26	2.10	0
84	1	40	1600	11	1	0	0	1	23,006	BLP	3.18	20.35	$4.4e^{-6}$
264	1	27	729	12	1	0	0	0	22,163	VERT	2.97	4.42	$1.6e^{-2}$
423	0	25	625	10	1	0	0	1	20,942	VERT	2.87	1.80	$3.5e^{-2}$
595	0	36	1296	7	1	0	0	1	20,781	VERT	2.83	12.89	$4.4e^{-2}$
38	1	42	1764	14	1	0	0	0	20,506	BLP	2.70	39.17	$7.8e^{-2}$

Weights and standardised residuals are computing using the Lin (2013) structured MM-estimator and the Mahalanobis distance is computed using the MCD estimator.

These 15 observations are the (only) removed/downweighted (i.e. weight < 0.1) in the Lin (2013) structured MM-estimator.

The first 13 observations as listed here, are also (in almost identical order) the top 13 highest earners (income 1978).

C.4 Covariate information observations with largest squared Mahalanobis distances

Table 11: Covariates 15 observations with the largest squared Mahalanobis distances (good leverage points)

id	T	Age	Age^2	Ed uc.	Bl ack	Hisp.	Mar ried	No deg.	Inc. 1978	Out lier	St. res.	Mah. dist.^{2}	w
453	0	55	3025	3	1	0	0	1	5,844	GLP	-0.11	351.45	1.00
559	0	54	2916	11	1	0	0	1	7,813	GLP	0.23	300.15	0.99
471	0	50	2500	10	0	1	0	1	0	GLP	-1.24	169.12	0.72
563	0	50	2500	8	1	0	1	1	8,997	GLP	0.56	162.80	0.94
236	1	49	2401	8	0	0	1	1	16,717	GLP	1.29	156.26	0.70
34	1	48	2304	4	1	0	0	1	6,552	GLP	0.36	124.39	0.97
75	1	46	2116	13	1	0	0	0	647	GLP	-0.57	82.46	0.94
41	1	46	2116	8	1	0	1	1	3,094	GLP	-0.44	80.31	0.96
72	1	45	2025	5	1	0	1	1	8,547	GLP	0.45	79.30	0.96
266	1	46	2116	8	1	0	0	1	0	GLP	-0.64	73.57	0.92
303	0	45	2025	11	1	0	0	1	11,796	GLP	1.19	62.71	0.74
571	0	45	2025	9	1	0	0	1	4,845	GLP	0.08	59.38	1.00
535	0	44	1936	9	1	0	1	1	12,359	GLP	1.29	53.41	0.70
60	1	41	1681	4	1	0	1	1	7,285	GLP	0.32	53.37	0.98
106	0	44	1936	11	1	0	0	1	0	GLP	-0.59	50.89	0.93

Weights and standardised residuals are computing using the Lin (2013) structured MM-estimator and the Mahalanobis distance is computed using the MCD estimator.