Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis BSc Econometrics and Economics

Outlier Robust Regression Discontinuity Designs

Francisco Magalhães Portilha
(479126)

Supervisor:	Jens Klooster
Second assessor:	Dr. Eoghan P. O'Neill
Date final version:	1st July 2024

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

Analysing the statistical properties of classical RDD estimation methods, and more robust RDD estimation (based on Huber and Tukey loss functions), when faced with outliers. A simulation with 6 outlier scenarios is performed, in which, for each outlier scenario a sample is generated $r = 10,000$ times, and, the 4 estimation methods are used to estimate the ATE. A power simulation and an asymptotic are also performed with $r = 1,000$, to gain insight of the power function and asymptotic properties.

Contents

1 Introduction

Regression Discontinuity Designs (RDD) are a useful tool for estimating average treatment effects (ATE) that were initially introduced by Thistlewaite and Campbell (1960) as "a method for testing causal hypotheses". In recent empirical literature, RDD's have been used extensively to estimate ATE's in various fields of research such as education (Angrist & Lavy, 1999; Taylor, 2014; Angrist & Rokkanen, 2015), finance (Chang et al., 2015; Dittmar et al., 2020), labour economics (Hijzen et al., 2017; Kreiner et al., 2020), and many others. RDD's exploit discontinuities in the data to estimate the ATE at that threshold. For example, Kreiner et al. (2020) exploit a discontinuity in the Danish youth minimum wage when turning 18 years old, to estimate ATE on youth employment.

In empirical settings, it is common practice to conduct some robustness and falsification checks to ensure that our estimates are robust. In the case of RDD a commonly used approach in empirical settings is to use a "donut" RDD (Barreca et al., 2011). This approach removes the observations close to the discontinuity in the running variable and is motivated by improper sorting around the threshold, or as proposed by Noack and Rothe (2023) due to the observed running variable and the "natural" running variable being different within a small strip at the threshold. Auerbach, Cai and Rafi (2024) show that under certain assumptions donut RDD can eliminate the effect of spillovers - when the outcome of an observation is affected by the outcome of its neighbours - at the threshold.

However, RDD estimates based on other robust estimators could be more accurate and efficient, if they do not remove the observations at the discontinuity. To investigate this, the class of M-estimators (Huber, 1964; Huber & Ronchetti, 2009) will be analysed within RDD, as this class has been shown to be more robust when the data contains outliers. This leads to the following research hypothesis: Are RDD estimates based on robust estimators from the class of M-estimator more robust than classical RDD estimates such as (local linear regressions RDD) when faced with outliers? A secondary research hypothesis also arises: How do M-estimator based RDD's compare to other robustness checks currently being used (donut RDD) in the presence of outliers?

Answering these research questions is of high importance as it can allow for more a robust estimation of treatment effects by RDD. To answer the research question, ATE will be estimated using RDD with the different estimation methods, to compare how more robust RDD estimations perform versus classical RDD estimation methods (local linear regression) (Imbens & Lemieux, 2008), and to other falsification methods currently used (donut RDD) (Noack & Rothe, 2023).

The research questions are tackled by a simulation study, in which different outliers scenarios are generated and RDD based on classical estimation methods will be compared to the RDD based on the class of M-estimators. The outline of this thesis is as follows, section 2 discusses the methodology for RDD, the set-up for the simulation is explained in section 3, followed by the main results in section 4, then a small application is performed in section 5, and lastly a conclusion.

2 Methodology

In this section the theoretical framework for sharp, donut, and robust RDD will be outlined. Followed by a description of the different estimation methods for RDD. For sharp RDD I follow the theoretical framework (and estimation methods) summarised by Imbens and Lemieux (2008); Cattaneo and Titiunik (2022), which analyses RDD from a Rubin Causal Model framework with potential outcomes (Rubin, 1974; Imbens & Rubin, 2015). Robust estimation is based on the literature on robust statistics from Hampel et al. (1986); Huber and Ronchetti (2009). And donut RDD is based on (Noack & Rothe, 2023).

2.1 Theoretical Framework

2.1.1 Potential Outcomes and Average Treatment Effects

Within the context of Rubin Causal Model the goal is estimate the causal effect of a treatment variable (binary intervention $T_i \in \{0, 1\}$). The notation used by Cattaneo and Titiunik (2022) is adopted and we have, for a sample of Y_i 's, $Y_i(0)$ denotes the outcome of observation i without treatment (if $T_i = 0$), and $Y_i(1)$ with treatment (if $T_i = 1$). The treatment effect can be calculated as $Y_i(1) - Y_i(0)$, however, the fundamental problem of causal inference is the inability to observe both the outcome of an observation having received treatment and without having received treatment (Imbens & Angrist, 1994; Imbens & Lemieux, 2008; Cattaneo & Titiunik, 2022). Therefore this must be estimated by $\mathbb{E}[Y_i(1) - Y_i(0)]$, and this is possible with a RDD. This also considers that the ATE might be heterogeneous across observations, by analysing the average treatment effect across the sample.

2.1.2 Regression Discontinuity Designs

In RDD, additionally to Y_i's and T_i's we also observe a running variable X_i's and may observe another covariate Z_i's which are both known to be unaffected by the treatment. Hence, for each observation $i = 1, 2, ..., n$ we observe (Y_i, T_i, X_i, Z_i).

The key feature of RDD is that the treatment is determined (completely or partly) by X_i being above, or below, a certain threshold c. RDD's leverage this to estimate average treatment effect. In sharp RDD the treatment is solely determined by the running variable X_i being above or below the threshold, $P(T_i = 0|X_i < c) = 1$ and $P(T_i = 1|X_i \geq c) = 1$.

For all 3 estimations methods described below the set-up is the same.

2.2 Estimation Methods

2.2.1 Local Linear Regression

The common approach to estimate sharp RDD is to use local linear regression (Linear RDD), which consist of fitting a linear regression (by OLS) on either side of the cutoff (Imbens & Lemieux, 2008). Alternatively, it is numerically equivalent to estimate with a single regression

(also including other possible covariates Z_i):

$$\min_{\alpha,\beta,\tau,\gamma} \sum_{i,c-h<X_i<c+h} (e_i)^2 \quad , e_i = e_i(\alpha,\beta,\tau,\gamma) = Y_i - \alpha - \beta(X_i - c) - \tau T_i - \gamma(X_i - c)T_i - \delta' Z_i \quad (1)$$

Where e_i is the i-th residual, and $\hat{\tau}$ is the estimated ATE. As recommended by Imbens and Lemieux (2008) the simple rectangular kernel will be used, as there isn't much improvement from using more sophisticated kernels unless if the estimate is highly sensitive to bandwidth choice, making the estimation questionable.

2.2.2 Donut RDD

Donut RDD uses similar estimation method as sharp RDD but eliminates the observation within a small strip around the threshold $|X_i - c| < d$:

$$\min_{\alpha,\beta,\tau,\gamma} \sum_{i,d<|X_i-c|<h} (e_i)^2 \quad (2)$$

With residuals as defined in (1). Like in Noack and Rothe (2023) a $\delta = 0.1$ is used through the rest of this study.

2.2.3 M-Estimates and Robust RDD

A Maximum-Likelihood Type Estimate (M-Estimates) is an estimate W_i that is defined by a minimization problem of the form: $\min \sum \rho(x_i; W_i)$; or by an implicit equation: $\sum \psi(x_i; W_i) = 0$, where $\psi(x_i;\theta) = \partial\rho(x_i;\theta)/\partial\theta$ (Hampel et al., 1986; Huber & Ronchetti, 2009). Often one is interested estimating location parameters, such that the form of the minimization problem becomes: $\min \sum \rho(x_i - W_i)$. Least-squares estimates are a special case where $\rho(e_i) = \frac{1}{2}e_i^2$.

A simple way to make estimates less sensitive to outliers (than OLS) is to choose a criterion function that grows slower than the squares of the residuals, then minimise its sum instead. One choice of criterion function is the least absolute deviation (LAD), which uses the absolute value function $\rho(e_i) = |e_i|$ as criterion. The advantage of the LAD is that all observation have equal weights, hence outliers will have less influence over the estimated parameters than in OLS. The disadvantage however is efficiency loss. To tackle this a loss function incorporating the benefits of OLS and LAD was proposed by Huber (1964):

$$\rho_H(e_i) = \begin{cases} \frac{1}{2}e_i^2 & \text{if} \quad |e_i| \le c \\ c|e_i| - \frac{1}{2}c^2 & \text{if} \quad |e_i| > c \end{cases} \quad (3)$$

With $c = 1.345$. Another choice of loss function is the Tukey loss function, which similarly to Huber's loss function in that it has a quadratic behaviour near zero, but weights observations away from 0 (outliers) even less than the LAD :

$$\rho_T(e_i) = \begin{cases} \frac{c^2}{6}\left(1 - \left[1 - \left(\frac{r}{c}\right)^2\right]^3\right) & \text{if} \quad |e_i| \le c \\ \frac{c^2}{6} & \text{if} \quad |e_i| > c \end{cases} \quad (4)$$

With $c = 4.685$. In practice most criterion functions are scale variant therefore we use the scaled criterion $\rho(\frac{e_i}{s})$, and the scale must be estimated first. Using this criterion function to estimate ATE with RDD:

$$\min_{\alpha,\beta,\tau,\gamma} \sum_{i,c-h<X_i<c+h} \rho\left(\frac{e_i}{s}\right) \tag{5}$$

Where the scale is estimated by an iterative procedure and with residuals as defined in (1). This should give more robust estimates when outliers are present than classical RDD estimation methods.

3 Simulation

To test the research hypotheses a simulation study has been conducted, consisting of 3 parts: base, power and asymptotic simulations. In which, a sample with a subgroup receiving a treatment, is generated, then the ATE are estimated based on the different RDD estimation methods (OLS, Robust Huber, Robust Tukey and Donut). This is replicated a number times for 6 different outlier scenarios to understand the statistical properties of the different RDD estimation methods when faced with and without outliers.

3.1 Data Generating Process

The simulation set-up will be partly based on the simulation from Noack and Rothe (2023), in which the running variable and the outcomes are estimated using the following DGP: $X_i \sim U(-1, 1)$, $T_i = \mathbf{1}(X_i > 0)$, $Y_i = \mu(X_i) + \varepsilon_i$, $\varepsilon_i \sim N(0, 0.5)$. [1] But, in contrast to Noack and Rothe, a linear outcome function with a homogeneous treatment is used to generate the outcomes:

$$\mu_{BL}(X_i) = \alpha + \beta X_i + \tau T_i \tag{6}$$

Where α and β are the parameters from a linear function and τ is the treatment effect - in this case the treatment effect is homogeneous for those who receive treatment as it does not depend on i. Since the interest of focus is the ATE, the parameters α and β are set at 0 and 1 respectively. In this simulation study two τ's are of special attention, namely -0.5 and 0. These values enable a clear demonstration the properties of the robust RDD estimation methods versus the classical RDD.

3.2 Simulation Procedure

The simulation study has 3 parts: First, the base simulations for the two τ's of special interest ($\tau = -0.5, 0$) use samples of size $N = 250$ and are simulated $R = 10,000$ times. Second, the power simulation of the power functions, for which simulation with 17 different values of τ in the range $[-2, 2]$ are performed with samples of size $N = 250$, but due to the increased computation time these are performed $R = 1,000$ times. Lastly, to understand the asymptotic properties of the different RDD estimators, a simulation is performed for 12 exponentially increasing sample

[1] Where $\mathbf{1}()$ is the indicator function.

sizes [2] from 60 to 10,000 for the two τ's of special interest, with the number of outliers staying constant. These are also performed $R = 1,000$ times due to the computation time.

3.3 Outlier Scenarios

The 6 different data contamination scenarios are the following: Scenario 1 does not contain any outliers, scenarios 2 and 3 contain 2 outliers, and, scenarios 4, 5 and 6 have 4 outliers. For scenario 2 both outliers are outside the right side of the "donut" strip, and, are positive and of moderate size. In scenario 3 the outliers are like in scenario 2 but of a large size. Scenario 4 has 2 large outliers like scenario 3 plus 2 large negative outliers on the left side outside of the "donut" strip. Scenario 5 has 4 outliers of the same type as scenario 4 but inside the "donut" strip. And, scenario 6 is similar to scenario 5 but there is one positive and one negative outlier on either side of the threshold. Figures A.1 and A.10 display one sample from each scenario with $\tau = -0.5$ and $\tau = 0$ respectively. The data contamination rates are 0.8% for scenario 2 and 3, and, 1.6% for the remaining scenarios (for $N = 250$).

Outliers are generated as follows: One or more observations are searched in the desired location of the running variable - either $[c - \delta, c]$ or $[c, c + \delta]$ for outliers inside the "donut" strip (where $\delta = 0.1$), or, $[c - 2\delta, c - \delta]$ or $[c + \delta, c + 2\delta]$ for the outliers outside the "donut" strip - and then the outcomes are changed to 3.5 for the moderate outliers (scenario 2), and either -10 or 10 for the remaining scenarios. If none or not enough observations in the desired location are found then then simulation proceeds with the ones that have been generated. However, the probability of this happening is very low as we expect 12.5 observations per interval were at most 2 are needed (for $N = 250$).

Using samples based these 6 different scenarios, ATE are estimated based on the 4 different RDD estimation methods (OLS, Robust Huber, Robust Tukey and Donut).

3.4 Evaluation Metrics

To evaluate the statistical properties of the different RDD estimation methods when faced with outliers, empirical evaluation metrics are calculated for each sample of ATE estimates (from the 4 estimation methods and 6 scenarios). The sample bias, standard deviation, skewness and kurtosis, root mean squared error, jarque-bera test, correct coverage and length of confidence intervals, type I and type II errors and power functions of t-tests.

Correct coverage of confidence intervals: $\frac{1}{R}\sum_{r=1}^{R} \mathbf{1}(lb_{r,\alpha} \leq \tau \leq ub_{r,\alpha})$, where $ub_{r,\alpha}$ and $lb_{r,\alpha}$ are the lower and upper bound of the confidence intervals respectively; size of confidence intervals: $\frac{1}{R}\sum_{r=1}^{R} ub_{r,\alpha} - lb_{r,\alpha}$; Type I error of t-test for $H_0(\tau) : \hat{\tau} = \tau$, versus a two-sided alternative: $\frac{1}{R}\sum_{r=1}^{R} \mathbf{1}(p < \alpha)$, where p is the p-value of the t-test ; Power functions $\pi(\tau)$ of the t-test for $H_0 : \tau = 0$ for $\tau = [-2, -1.75, ..., 1.75, 2]$. Significance levels (α) of .1, .05 and .01 where used for the t-tests, the correct coverage and size of the c.i., and .05 for the power functions.

[2] $N = 250 * 1.6^i, i = -3, ..., 8$

4 Results

4.1 Single Sample Preview

Figure A.1 displays the fitted linear regressions based on the 4 different RDD estimation methods (OLS, Robust Huber, Robust Tukey and Donut), along with the estimated ATE and corresponding significance levels, for the 6 outlier scenarios from the simulation with $\tau = -0.5$. It can be observed that already in the case with moderate outliers (scenario 2), the size and the significance of the OLS and Donut estimates seem to be affected by the outliers, as they are already unable to reject that the ATE are significantly different from zero, unlike the robust counterparts. This suggests that less than 1% of data contamination of moderate size is already enough to affect the size (and significance) of the ATE estimates. In scenario 3, the two large outliers are enough to shift the sign of the OLS and Donut estimates to positive (althought not significant). In the extreme scenario 4, OLS and Donut both estimate a positive ATE significant at 5%, whilst the Robust Huber and Tukey still show significant negative ATE. Scenario 5 suggests that if the outliers are within the "donut" strip the Donut estimate is now correctly negative and significant, but the OLS estimate seems even more affected by the outliers. And if the outliers are symmetric (scenario 6) it seems like OLS is now correctly estimating the ATE, however it is still unable to reject that the ATE are significantly different from zero.

This simple one sample preview, already shows a suspicious behaviour of the classical RDD estimation methods when face with outliers. The OLS and Donut estimates vary between -0.6 and 1 within these six samples, due to just 1.6% of data contamination. Whilst the Huber and Tukey show a much more stable behaviour between -0.6 and -0.3. The robust estimation methods also don't seem to have the significance of their estimates (as) disturbed by the outliers.

The single sample preview from the simulation with $\tau = 0$ (Figure A.10), also increase the suspicion of unreliability in the classical estimation methods. The OLS and Donut estimation methods display a similar incoherence as in the simulation with $\tau = -0.5$. Of special attention is scenario 4, for which OLS and Donut would incorrectly estimate a positive ATE, significant at 5%, while the Huber and Tukey still estimate an ATE quite close to zero, and, would not incorrectly reject that the ATE are nonexistent.

4.2 Point Estimation

To analyse the properties of point estimates from the different RDD estimation methods, when faced with outliers, the histograms of the 10,000 ATE estimated with each of the 4 methods are plotted in Figure 1, for each scenario (from the simulation with $\tau = -0.5$). The exact sample bias, standard deviation, root mean squared error, skewness, kurtosis, and Jarque-Bera p-values are displayed in Tables A.1 and A.2.

As expected from theory, in scenario 1 (no outliers), all estimation methods are unbiased. As unbiased estimators, OLS (0.127) is the most efficient, closely followed by Huber (0.130) and Tukey (0.131). [3] The Donut (0.156) is notably less efficient due to the smaller effective sample size it's estimated on. With just two moderate outliers (scenario 2), all estimation methods,

[3] Theses values are the standard deviation of the ATE estimates, Table A.1.

except for Tukey, are now biased. Huber (0.042) is only slightly biased, while OLS (0.193) and Donut (0.296) are a bit more.[4] This confirms the initial suspicion from the single sample preview, that less than 1% of data contamination of moderate size is enough to affect the point estimate of the classical RDD estimates. These moderate outliers also increase the standard deviation of the estimates, from all 4 methods, by 1.5% to 3% compared to scenario 1 with Donut being the most affected.

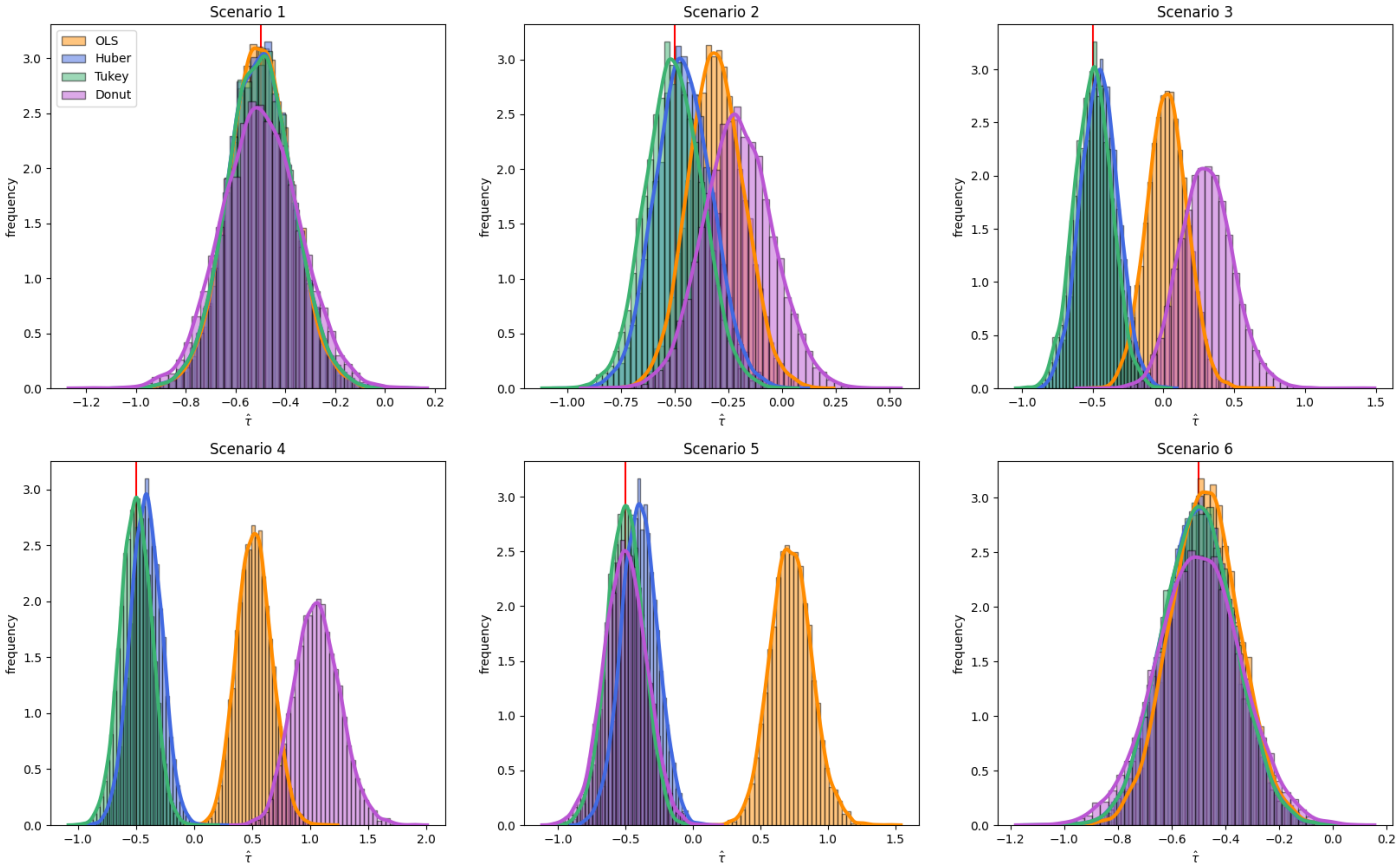

Figure 1: Histograms of the estimated ATE based on the 4 different RDD estimation methods, for each of the (outliers) scenarios. From the simulation with $r = 10,000$, $\tau = -0.5$.

If the size of the two outliers is increased by a factor of 2.86x (from 3.5 in scenario 2 to 10 in scenario 3), then the bias of the OLS and Donut estimates become 2.7x larger, displaying the sensitivity of the classical estimators to the size of the outliers. Additionally, if the amount of data contamination is doubled (from 0.8% in scenario 3 to 1.6% in scenario 4) then the bias of the OLS, Donut and Huber estimates also (almost) double, displaying the sensitivity, not only of the classical estimates but also of the robust estimator with convex loss function, to the amount of contaminated data. Although the influence on the bias of the OLS and Donut estimates seems similar whether the size of outliers or the amount of contamination is increased, not the same can be said for the standard deviations, which increase by 10% and 17% from scenario 2 to 3, but only increase by 4% and 5% from scenario 3 to 4, for OLS and Donut respectively. Hence, despite the similar influence on the bias of the point estimates from increasing the size of outliers or the amount of contamination, the latter is "anchoring" the biased estimates, while

[4]These values are the bias of the ATE estimates, Table A.1.

the former is making the estimates more unstable.

Overall, the robust RDD estimation methods still seem very accurate despite the outliers. The opposite is true for OLS and Donut, with special attention to the Donut being even more biased than the OLS for all scenarios, except scenarios 5 and 6 where the outliers are inside the donut strip, but even in this case the Tukey estimator still has a lower RMSE. Tukey seems to perform better overall in the presence of outliers since it has the lowest RMSE in the 4 scenarios with asymmetric outliers (scenarios 2, 3, 4 and 5). As expected from theory, in scenario 1 OLS has the lowest RMSE, but also for scenario 6 (symmetric outliers). This could suggest that the OLS estimate is the best in this case, but with regards to interval estimation and inference this will prove to be completely false. Note that even if Tukey does not seem to be biased due to the outliers but it does increase the standard deviations of its estimates.

4.3 Interval Estimation

The percentage of confidence intervals with correct coverage, and their lengths, estimated with the 4 different RDD methods, from the simulation with $\tau = -0.5$, are reported in Table A.3. In scenario 1, as expected, all methods seem to have a coverage close to 95%, with OLS (0.502) having the smallest length closely followed by Huber (0.512) and Tukey (0.513). The confidence intervals estimated with the Donut (0.621) have a larger length, due to the smaller effective sample size it's estimated on.

Similar to the results from the point estimation, just 2 moderate outliers (scenario 2) are enough to affect the accuracy of our interval estimates obtained from OLS and Donut RDD's, as the correct coverage falls to 80% and 70% respectively, and the length increases by 20% for both methods. The correct coverage from OLS and Donut interval estimates then falls to 49% and 24% in scenario 3 and 0.6% and 0% in scenario 4, and the lengths increase by 75% followed by 30% for both methods. The correct coverage from the Tukey confidence intervals remains very close to 95% in all scenarios and the lengths are very stable, making this also the preferred estimator for interval estimation. The Huber, as for point estimation, is slightly more affected than Tukey but still performs very well, only falling as far as 88% correct coverage.

Notable that the correct coverage of all methods (except Donut) drops from scenario 4 to 5 as the leverage of the outliers increases, but the lengths stay the same since there is no increase in the size of the outliers or the amount of data contamination.

The results from the simulation with $\tau = 0$ are very similar. This again shows the sensitivity of estimation of ATE based on the classical RDD methods to outliers (now with regards to intervals estimation), when compared to the robust counterparts which seem much more stable.

4.4 Inference

The type I and type II errors from the different RDD estimation methods, from the simulation with $\tau = -0.5$, are displayed in Table A.4. The type I errors are naturally related to the correct coverage, since if the confidence interval does not cover the true parameter then the true null hypothesis will be rejected. Therefore, type II errors (for incorrect $h_0 : \hat{\tau} = 0$) are of focus in this section. As expected OLS (2.5%) has the smallest type II error in the case without outliers, closely followed by the Huber (3.1%) and Tukey (3.2%). With the Donut (11.5%) having quite a

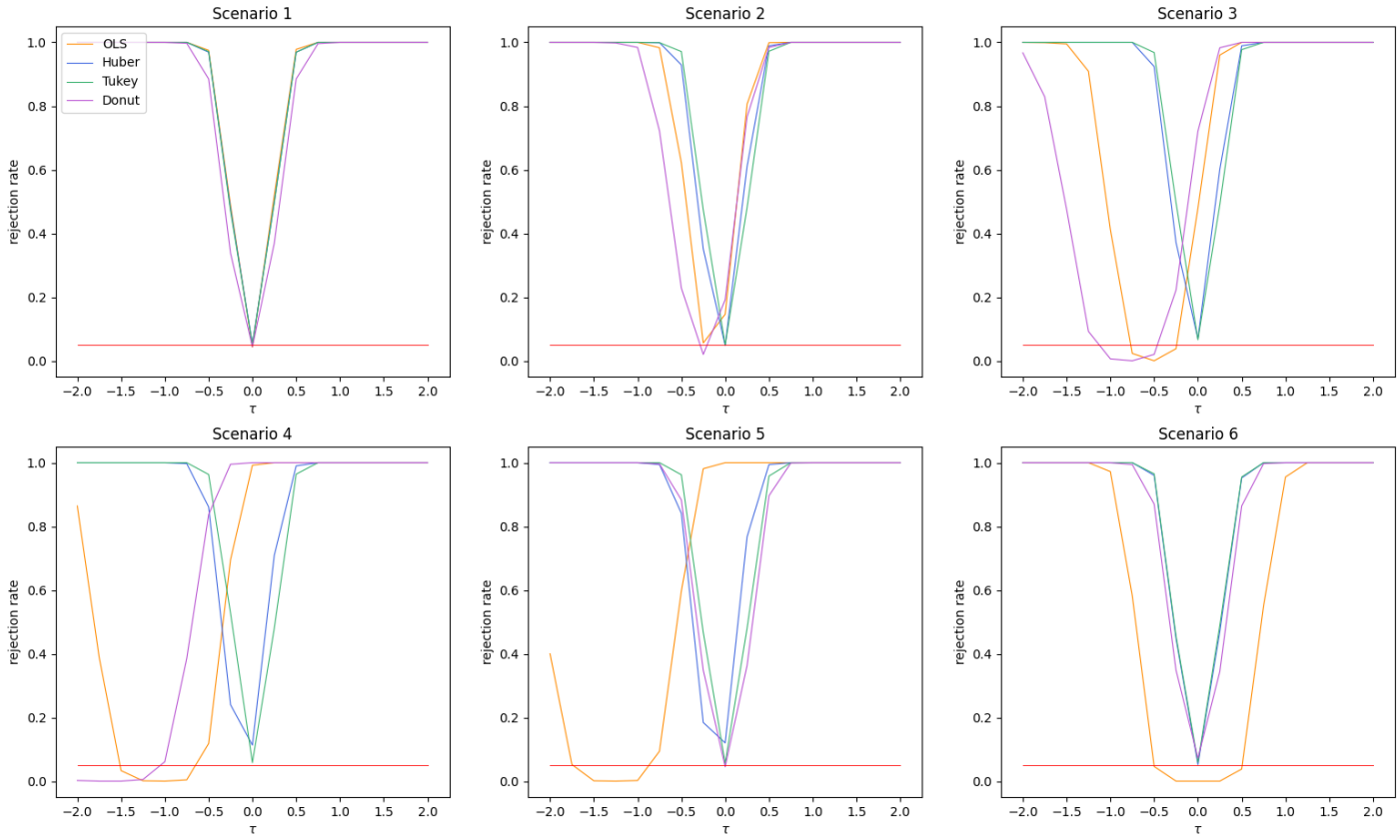

Figure 2: Power functions ($\pi(\tau)$) of the t-test for $H_0 : \tau = 0$, at a 5% significance level, of the different RDD estimation methods in each scenario. Simulation with $r = 100, \tau = [-2, -1.75, ..., 1.75, 2]$.

larger type II error. It can be observed that the type II error from the OLS and Donut estimates increases noticeably in scenarios 2 and 3, whilst the robust RDD estimates do not see their type II error affected. Notable that, in scenario 4 and 5 the Donut estimates, and in scenario 5 the OLS estimates, seem to have a low type II error but these estimates are now rejecting an incorrect null, but for the wrong reason, as the estimates are positive in these scenarios.

More broadly, the inference results from the power simulation are displayed in Figure 2. The power of the test is related to the type II error, therefore similar conclusions can be drawn for the level of $\tau = -0.5$. It can be observed that for other levels of τ in the range $[-2, 2]$, inference on $h_0 : \hat{\tau} = 0$ based on the classical RDD estimation methods is very affected by outliers. for the level of $\tau = 0$ it can be observed that these methods incorrectly reject this null (almost) 100% in scenarios 4 and 5. In these scenarios, the minimum of the power function is shifted to the left, and in the case the true parameter was of as large as 1.5, the classical estimation methods would only be able to reject the incorrect null less than 5% of the times.

In the special case of scenario 6 although OLS might have seemed appropriate for point estimation, the power of a t-test based on this method would still be greatly affected by the outliers. As the OLS power function is less than the power functions based on the robust RDD for all values of τ.

4.5 Asymptotic Analysis

The results from the asymptotic simulation are displayed in Figures A.2 - A.9 and Figures A.12 - A.18. It can be observed, as expected that the unreliability of the classical RDD estimation methods is no longer a concern as the sample size increases (if the absolute number of outlier remains constant for each scenario). However, for the 4 scenarios with asymmetric outliers (scenarios 2, 3, 4 and 5), not even with a sample of size $n = 10,000$ (making the amount of data contamination 0.02% and 0.04%) does the bias of the OLS and Donut estimates become as small as the Huber and Tukey. The standard deviation are all relatively close, and if we assume the bias to be negligible after a large n, the efficiency figures display that Donut is clearly less efficient overall, even when the outliers are within the "donut" strip, and OLS is more efficient after a large enough sample size.

The asymptotic properties of the interval estimations reveal a similar conclusion, with an enough large sample size the classical estimations are not unreliable anymore. The correct coverage tends towards the expected level, however, a very large sample size is required for correct interval estimation, in scenarios 4 and 5 OLS can only recover correct coverage of 80% when the sample is 10,000. Inference properties based on the classical RDD estimation methods also tend to regain the desired levels, but again, a very large sample size is required. The Huber and Tukey RDD's are very reliable from samples as large as 100, making them the preferred option when faced with outliers as long as the sample size is a restriction.

5 Application

To demonstrate the robust RDD estimation methods in a real-life setting, the study by Barreca et al. (2011), which started the "donut" RDD, has been revisited. In this paper, the authors reanalyse the results from Almond et al. (2010), who were interested in measuring the benefits from medical expenditures for at-risk newborns. To this end, they exploit a discontinuity in the assignment of medical treatments for newborns at a birth weight of 1500g [5], and found significant negative effects on mortality rates from being born just under the threshold, even with a general decrease in mortality rates as birth weight increases. Supporting the claim that medical care does have positive effects for at-risk newborns.

However, the study was subject to large heaps at the 100-gram and 1-oz multiples - which can be due to rounding or measuring constrains (Barreca et al., 2011) - of special attention was the heap at the threshold (1500g), which would render the RDD invalid if caused by manipulation. This was considered by Almond et al., by testing for manipulation of the running variable at the threshold and no evidence supporting this was found [6], hence, the results were deemed to be valid by the authors. On the other hand, Barreca et al. believe that, regardless of the absence in manipulation, the heap at the threshold is an outlier, since the observed mortality rates for this group are much higher than for those on either side of the cutoff. This led them to develop the "donut" RDD, to verify if the results were robust to this outlier (and possible nonrandom sorting at the cutoff). Using this method they concluded that the initial estimates were not robust. By simply removing the observation at the threshold the estimate was cut by half, and, with a "donut" of 3g they were already unable to reject that the ATE were significantly different

[5] Babies that are born with less than 1500g are called at-risk and receive extra care.
[6] This was tested for using McCrary (2008) density test for manipulation of the running variable in RDD.

from zero.

The concern about the validity of the estimates due to the outliers, makes this study a good example to demonstrate the robust RDD estimation methods. And, the need for a correct understanding, of the effects from medical care, for policy making, is of the utmost importance. Therefore, with these aims in mind, the "birth cohort linked birth-infant death data files" from the National Center for Health and Statistics from 1983-1991 and 1995-2002, which were used by Almond et al. (2010) and Barreca et al. (2011), were collected. The files consist of information from around 66 million US births certificates, and, linked information from death certificates if the newborn perished in the first year of life. Of these, more than 200,000 are within the optimal bandwidth determined by Almond et al. (85g on either side of the threshold), which was also used by Barreca et al..

When performing the analysis I realised that, in the case of binary dependent variable, OLS remains unbiased Heij et al. (2004), however this may not be the case for the robust estimation methods used. In this example as most of the observations are 0 (for not dead), and, when scaled the observation with 1 (for dead) are greater than the c used in both robust estimation methods, hence the ATE are estimated at 0 since it is more costly to give any other estimate. Therefore standard RDD estimation methods (OLS) can be used in RDD with binary dependent variables without further adjustments, but additional methodology about the correct way to estimate a binary dependent variable with robust methods is needed. [7]

6 Conclusion

The results from the simulation reveal very promising capabilities of RDD's based on the robust estimation methods from the class of M-estimators when faced with outliers. The Tukey estimator seems to perform the best overall across the outliers scenarios explored even when faced, and some (extra) sensitivity scenarios revealed that it remains quite robust even with more data contamination. The classical RDD estimation methods on the other hand, are very unreliable in the scenarios explored and would lead to misleading conclusions. Of special attention are scenarios when the true parameter is negative, whilst the classical estimation methods might report significant positive ATE (or vice-versa), and in scenarios when ATE is zero the classical estimation methods might report significant positive (or negative) results. These cases can lead to policy decisions that will have undesired effects. This can be prevented by estimating RDD based on Huber and Tukey loss functions.

We can confidently answer the main research with a positive note on robust RDD, but a methodology is necessary for the evaluation of RDD based on these robust methods setting with discrete dependent variables. Asymmetric loss functions can also be of interest, this can be of particular interest if one has an idea of the direction of the outliers, for instance if there is improper sorting at the threshold we can expect to be of large positive outliers on one side of the cutoff and large negative outliers on the other side (for a DGP with outcome 0 at the threshold). with regards to other falsification checks (Donut RDD) it was found that if the "donut" strip is misspecified then this method performs worse than regular OLS. In order to avoid the risk

[7] This was briefly explored but it was not implemented.

of misspecifying the "donut" size, the proposed robust RDD estimates can be used, which were shown to be more effective, even when the "donut" strip is indeed correctly specified.

Acknowledgements

I want to thank Jens Klooster, my supervisor, for suggesting this topic. I've enjoyed doing this research very much and that is partly thanks to having picked a nice topic, and the useful tips received, so for that I am thankful.

References

Almond, D., Doyle, J. J., Kowalski, A. E. & Williams, H. (2010). Estimating marginal returns to medical care: Evidence from at-risk newborns. *The Quarterly Journal of Economics*, *125*(2), 591–634. doi: https://doi.org/10.1162/qjec.2010.125.2.591

Angrist, J. D. & Lavy, V. (1999). Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, *114*(2), 533–575. doi: https://doi.org/10.1162/003355399556061

Angrist, J. D. & Rokkanen, M. (2015). Wanna get away? regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*, *110*(512), 1331-1344. doi: https://doi.org/10.1080/01621459.2015.1012259

Auerbach, E., Cai, Y. & Rafi, A. (2024). *Regression discontinuity design with spillovers.* doi: https://doi.org/10.48550/arXiv.2404.06471

Barreca, A. I., Guldi, M., Lindo, J. M. & Waddell, G. R. (2011). Saving babies? revisiting the effect of very low birth weight classification. *The Quarterly Journal of Economics*, *126*(4), 2117–2123. doi: https://doi.org/10.1093/qje/qjr042

Cattaneo, M. D. & Titiunik, R. (2022). Regression discontinuity designs. *Annual Review of Economics*, *14*, 821–851. doi: https://doi.org/10.1146/annurev-economics-051520-021409

Chang, Y.-C., Hong, H. & Liskovich, I. (2015). Regression discontinuity and the price effects of stock market indexing. *The Review of Financial Studies*, *28*(1), 212–246. doi: https://doi.org/10.1093/rfs/hhu041

Dittmar, A., Duchin, R. & Zhang, S. (2020). The timing and consequences of seasoned equity offerings: A regression discontinuity approach. *Journal of Financial Economics*, *138*(1), 254-276. doi: https://doi.org/10.1016/j.jfineco.2020.04.017

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions.* Wiley. doi: http://dx.doi.org/10.1002/9781118186435

Heij, C., de Boer, P., Franses, P. H., Kloek, T. & van Dijk, H. K. (2004). *Econometric methods with applications in business and economics.* Oxford University Press. doi: https://doi.org/10.1093/oso/9780199268016.001.0001

Hijzen, A., Mondauto, L. & Scarpetta, S. (2017). The impact of employment protection on temporary employment: Evidence from a regression discontinuity design. *Labour Economics*, *46*, 64-76. doi: https://doi.org/10.1016/j.labeco.2017.01.002

Huber, P. J. (1964). Robust estimation of location parameter. *Annals of Mathematical Statistics*, *35*(1), 73-101. doi: https://doi.org/10.1214/aoms/1177703732

Huber, P. J. & Ronchetti, E. M. (2009). *Robust statistics.* Wiley. doi: https://onlinelibrary.wiley.com/doi/book/10.1002/9780470434697

Imbens, G. W. & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467-475. doi: https://doi-org.eur.idm.oclc.org/10.2307/2951620

Imbens, G. W. & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615–635. doi: https://doi.org/10.1016/j.jeconom.2007.05.001

Imbens, G. W. & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences*. Cambridge University Press. doi: https://doi.org/10.1017/CBO9781139025751

Kreiner, C. T., Reck, D. & Skov, P. E. (2020). Do lower minimum wages for young workers raise their employment? evidence from a danish discontinuity. *The Review of Economics and Statistics*, *102*(2), 339–354. doi: https://doi.org/10.1162/rest_a_00825

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, *142*(2), 698-714. doi: https://doi.org/10.1016/j.jeconom.2007.05.005

Noack, C. & Rothe, C. (2023). *Donut regression discontinuity designs.* doi: https://doi.org/10.48550/arXiv.2308.14464

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. doi: https://psycnet.apa.org/doi/10.1037/h0037350

Taylor, E. (2014). Spending more of the school day in math class: Evidence from a regression discontinuity in middle school. *Journal of Public Economics*, *117*, 162-181. doi: https://doi.org/10.1016/j.jpubeco.2014.06.002

Thistlewaite, D. L. & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, *51*(6), 309–317. doi: https://doi.org/10.1037/h0044319

A Appendix

A.1 Results from simulation with $\tau = -0.5$

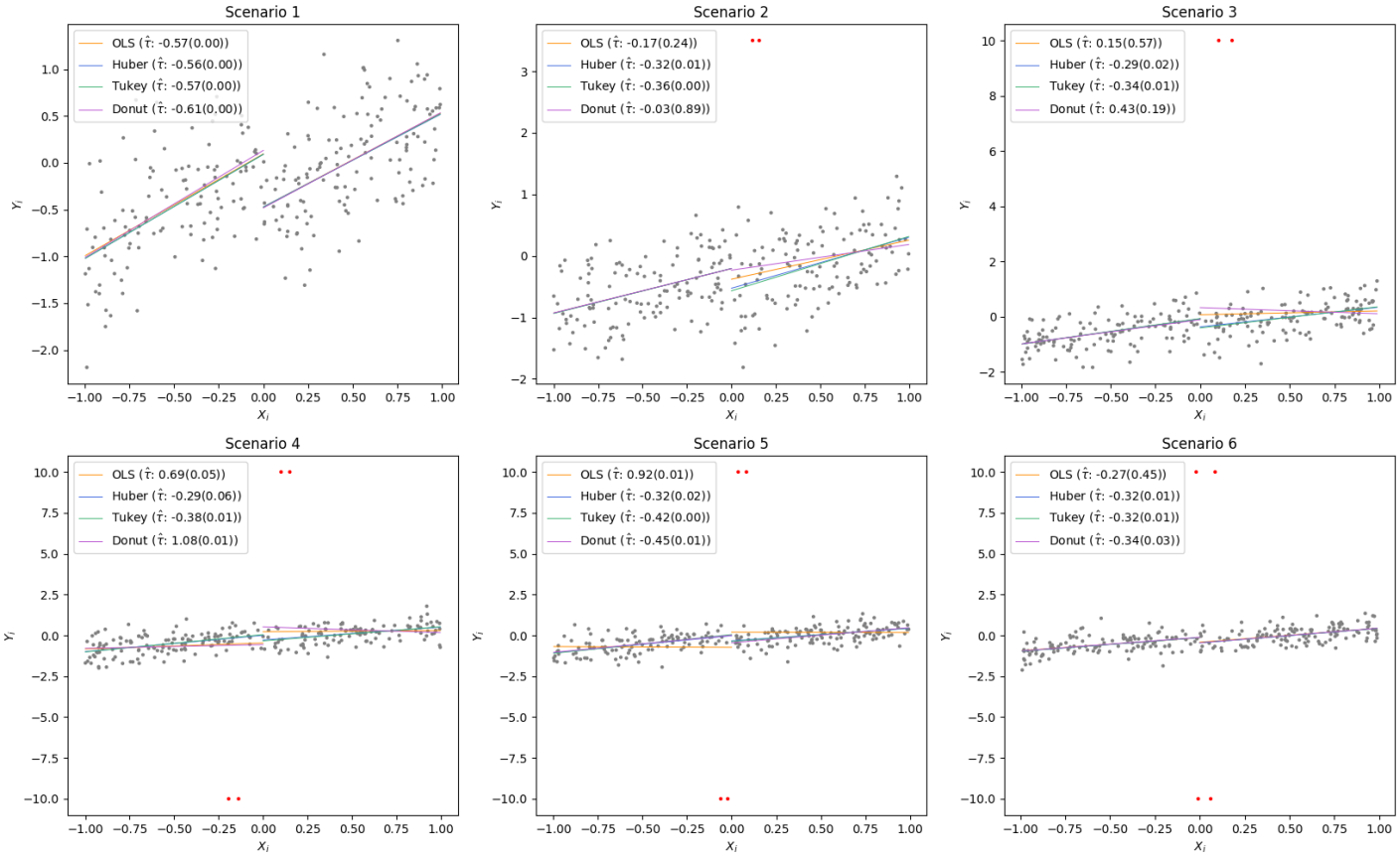

Figure A.1: The 6 different types of samples generated for each of the (outliers) scenarios with the fitted linear regressions estimated with the 4 different methods. The estimated ATE are also reported with their significance's. These are the first samples of each scenario from the simulation with $\tau = -0.5$.

Table A.1: Bias, standard deviation and root mean squared error of the point estimates of the treatment effect. From the simulation with $r = 10,000$, $\tau = -0.5$.

	Scenario 1			Scenario 2			Scenario 3		
	Bias	St.Dev.	RMSE	Bias	St.Dev.	RMSE	Bias	St.Dev.	RMSE
OLS	-0.001	0.127	0.127	0.193	0.129	0.232	0.526	0.142	0.545
Huber	-0.001	0.130	0.130	0.042	0.133	0.139	0.047	0.132	0.140
Tukey	-0.001	0.131	0.131	-0.001	0.133	0.133	0.004	0.132	0.132
Donut	0.000	0.156	0.156	0.296	0.161	0.337	0.802	0.189	0.824
	Scenario 4			Scenario 5			Scenario 6		
	Bias	St.Dev.	RMSE	Bias	St.Dev.	RMSE	Bias	St.Dev.	RMSE
OLS	1.020	0.148	1.031	1.228	0.154	1.238	0.024	0.128	0.130
Huber	0.087	0.134	0.160	0.103	0.134	0.169	-0.000	0.135	0.135
Tukey	0.001	0.134	0.134	-0.002	0.134	0.134	-0.000	0.135	0.135
Donut	1.561	0.199	1.574	-0.002	0.158	0.158	0.001	0.159	0.159

Table A.2: Skewness, kurtosis and p-value of the jarque-bera statistic of the estimated treatment effects. From the simulation with $r = 10,000$, $\tau = -0.5$.

	Scenario 1			Scenario 2			Scenario 3		
	Skew	Kurt	JB	Skew	Kurt	JB	Skew	Kurt	JB
OLS	0.022	3.117	0.038	0.078	3.017	0.006	0.072	3.035	0.010
Huber	0.028	3.111	0.042	0.053	3.043	0.064	0.015	2.982	0.767
Tukey	0.027	3.108	0.049	0.046	3.039	0.127	0.010	2.980	0.847
Donut	0.007	3.075	0.299	0.053	3.074	0.030	0.147	3.232	0.000
	Scenario 4			Scenario 5			Scenario 6		
	Skew	Kurt	JB	Skew	Kurt	JB	Skew	Kurt	JB
OLS	0.119	2.972	0.000	0.171	3.181	0.000	0.026	2.946	0.312
Huber	-0.001	3.025	0.873	0.010	2.983	0.861	0.016	2.939	0.374
Tukey	-0.010	3.029	0.767	-0.010	2.970	0.769	0.019	2.934	0.303
Donut	0.186	3.065	0.000	-0.003	3.007	0.982	-0.014	3.065	0.349

Table A.3: Correct coverage of the confidence intervals for τ and their length. For significance level of $\alpha = 0.05$. From the simulation with $r = 10,000$, $\tau = -0.5$.

	Scenario 1		Scenario 2		Scenario 3	
	C.C.	Length	C.C.	Length	C.C.	Length
OLS	0.950	0.502	0.800	0.606	0.491	1.046
Huber	0.948	0.512	0.936	0.520	0.938	0.520
Tukey	0.948	0.513	0.948	0.515	0.946	0.515
Donut	0.955	0.621	0.701	0.759	0.236	1.335
	Scenario 4		Scenario 5		Scenario 6	
	C.C.	Length	C.C.	Length	C.C.	Length
OLS	0.006	1.363	0.000	1.365	1.000	1.375
Huber	0.904	0.527	0.879	0.528	0.949	0.526
Tukey	0.946	0.516	0.944	0.517	0.944	0.515
Donut	0.000	1.748	0.949	0.620	0.947	0.619

Table A.4: Type I and type II errors of t-test for $h_0 : \hat{\tau} = \tau$ and $h_0 : \hat{\tau} = 0$ repectively, from the estimated treatment effects. From the simulation with $r = 10,000$, $\tau = -0.5$. For significance level of $\alpha = 0.05$.

	Scenario 1		Scenario 2		Scenario 3	
	T.I	T.II	T.I	T.II	T.I	T.II
OLS	0.050	0.025	0.200	0.480	0.509	1.000
Huber	0.052	0.031	0.064	0.070	0.062	0.074
Tukey	0.052	0.032	0.052	0.035	0.054	0.035
Donut	0.045	0.115	0.299	0.862	0.764	0.973
	Scenario 4		Scenario 5		Scenario 6	
	T.I	T.II	T.I	T.II	T.I	T.II
OLS	0.994	0.872	1.000	0.385	0.000	0.949
Huber	0.096	0.139	0.121	0.165	0.051	0.042
Tukey	0.054	0.035	0.056	0.038	0.056	0.038
Donut	1.000	0.158	0.051	0.117	0.053	0.116

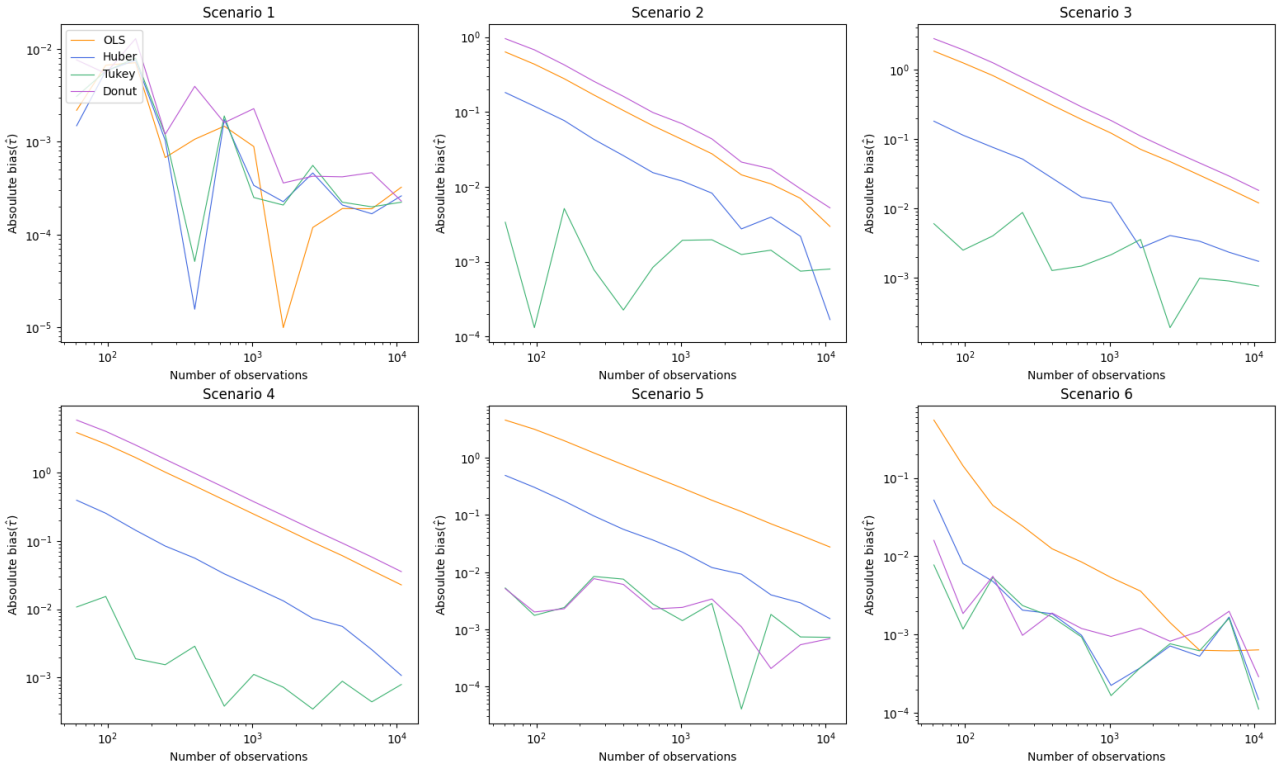

Figure A.2: Absolute value of the bias from the ATE estimates from the 4 estimation methods for increasing sample sizes. From the simulation with $r = 1,000$, $\tau = -0.5$.

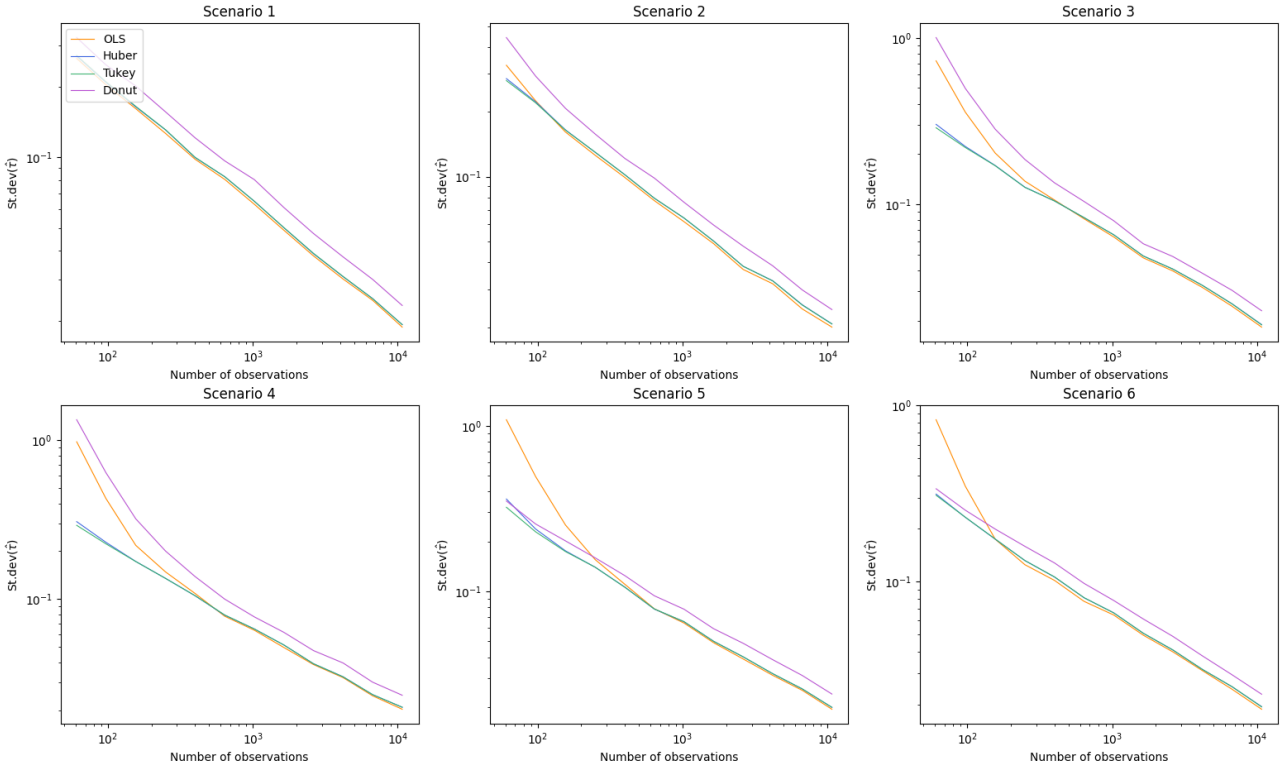

Figure A.3: Standard deviation of the ATE estimates from the 4 estimation methods for increasing sample sizes. From the simulation with $r = 1,000$, $\tau = -0.5$.

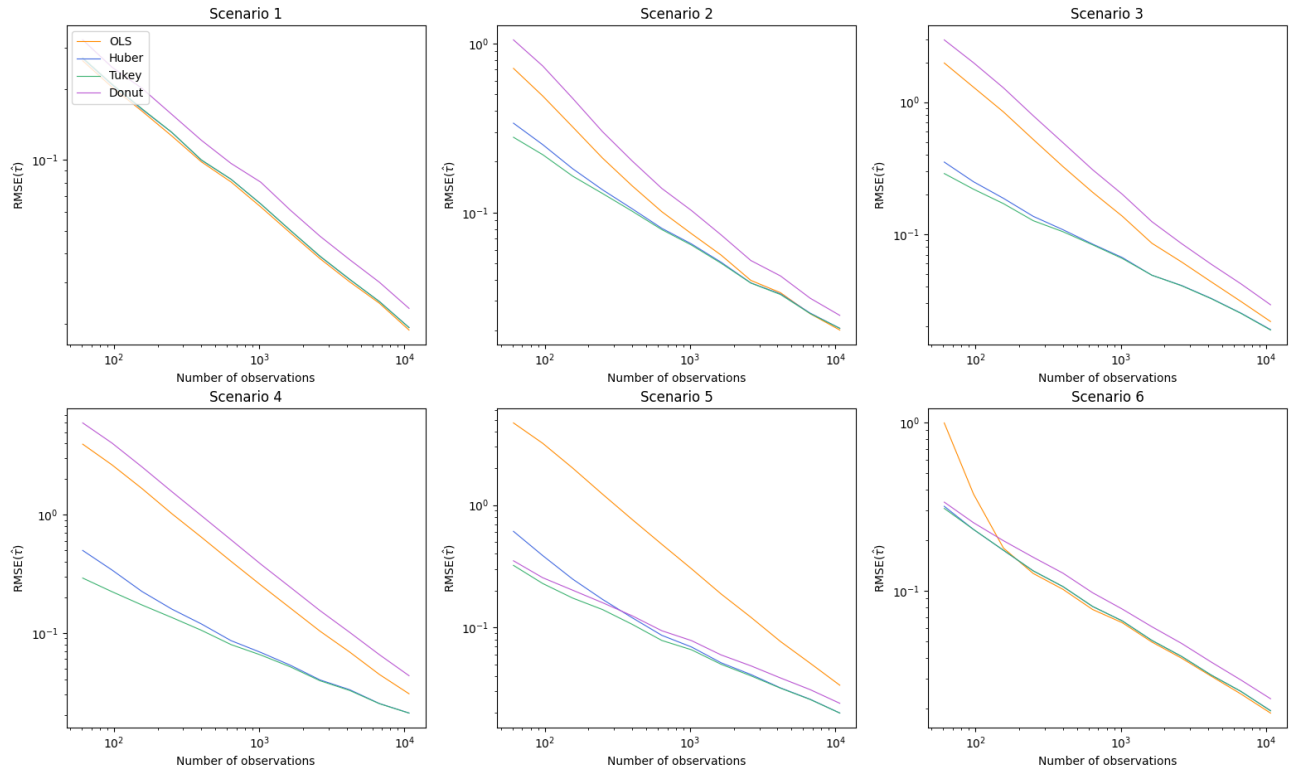

Figure A.4: RMSE of the ATE estimates from the 4 estimation methods for increasing sample sizes. From the simulation with $r = 1,000$, $\tau = -0.5$.

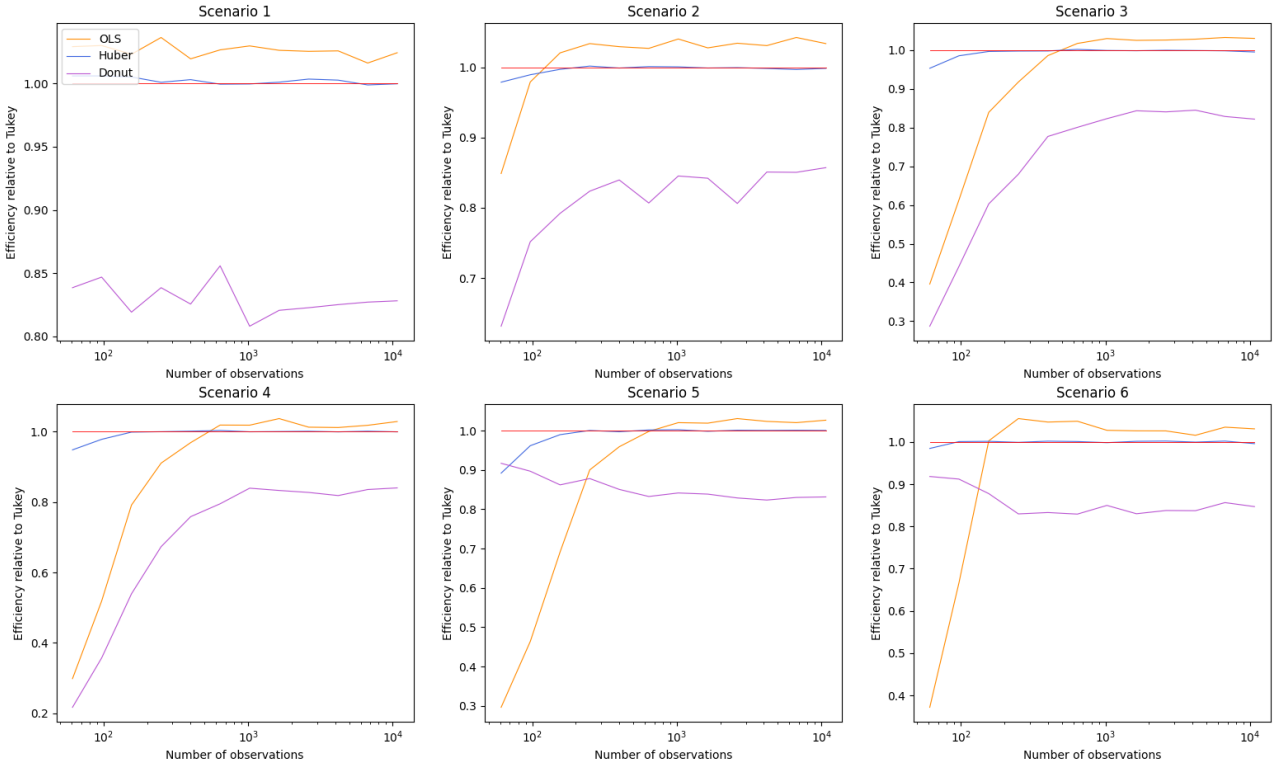

Figure A.5: Efficiency of the ATE estimates from the 3 estimation methods relative to Tukey estimates, for increasing sample sizes. Values above 1 mean that the estimate is more efficient than Tukey's. From the simulation with $r = 1,000$, $\tau = -0.5$.

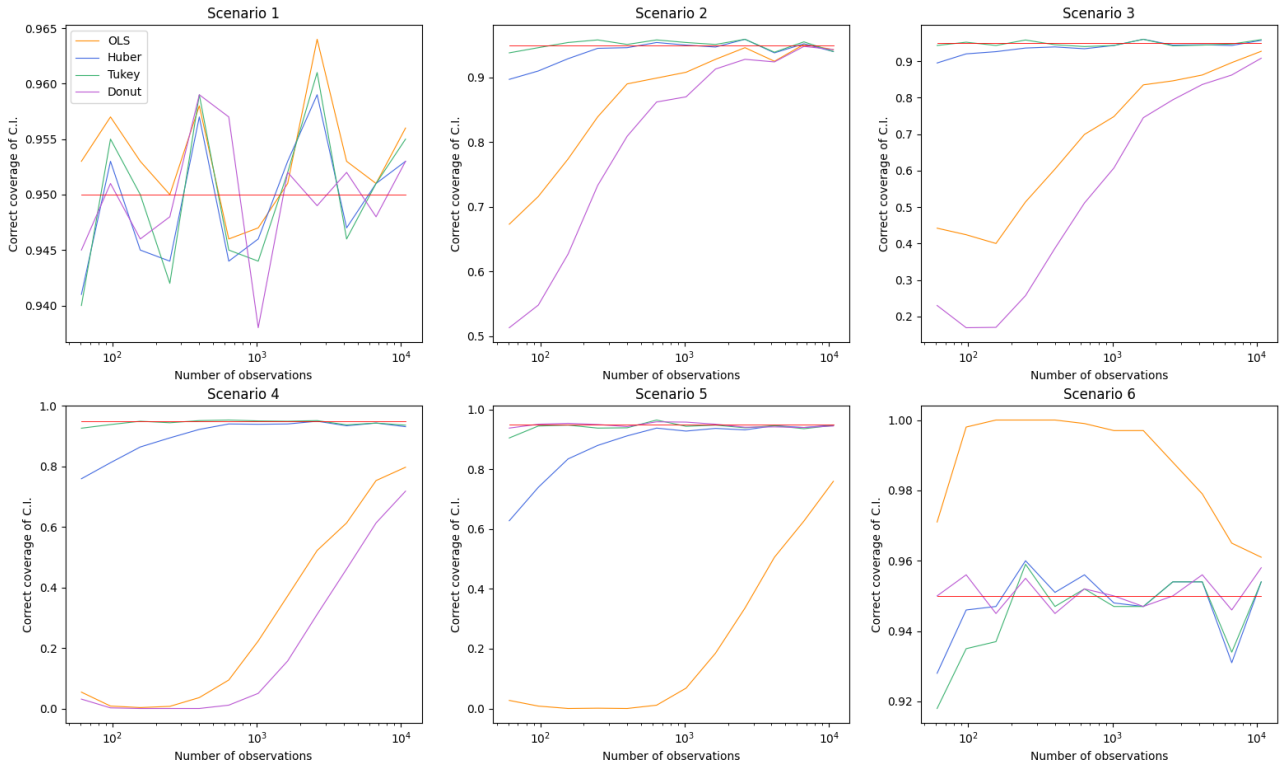

Figure A.6: Correct coverage of the confidence intervals of the ATE estimates from the 4 estimation methods, for increasing sample sizes. From the simulation with $r = 1,000$, $\tau = -0.5$.

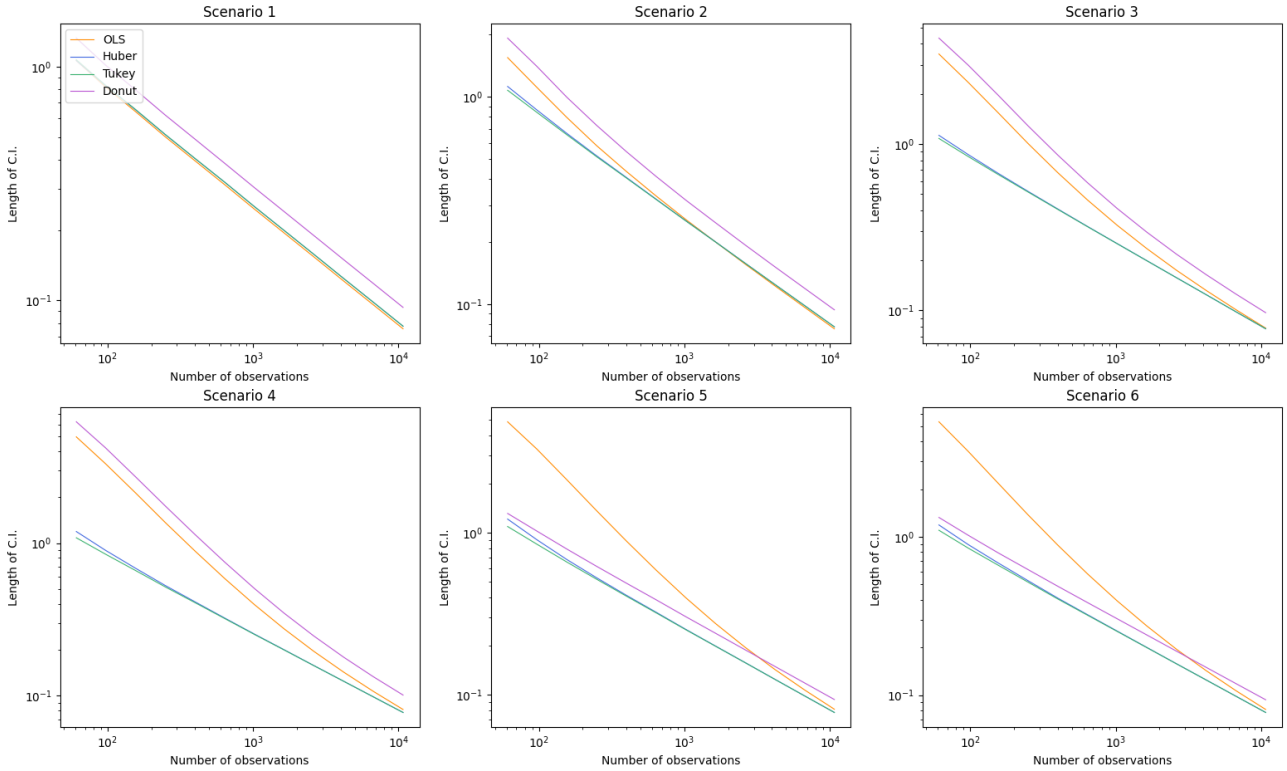

Figure A.7: Length of the confidence intervals of the ATE estimates from the 4 estimation methods, for increasing sample sizes. From the simulation with $r = 1,000$, $\tau = -0.5$.

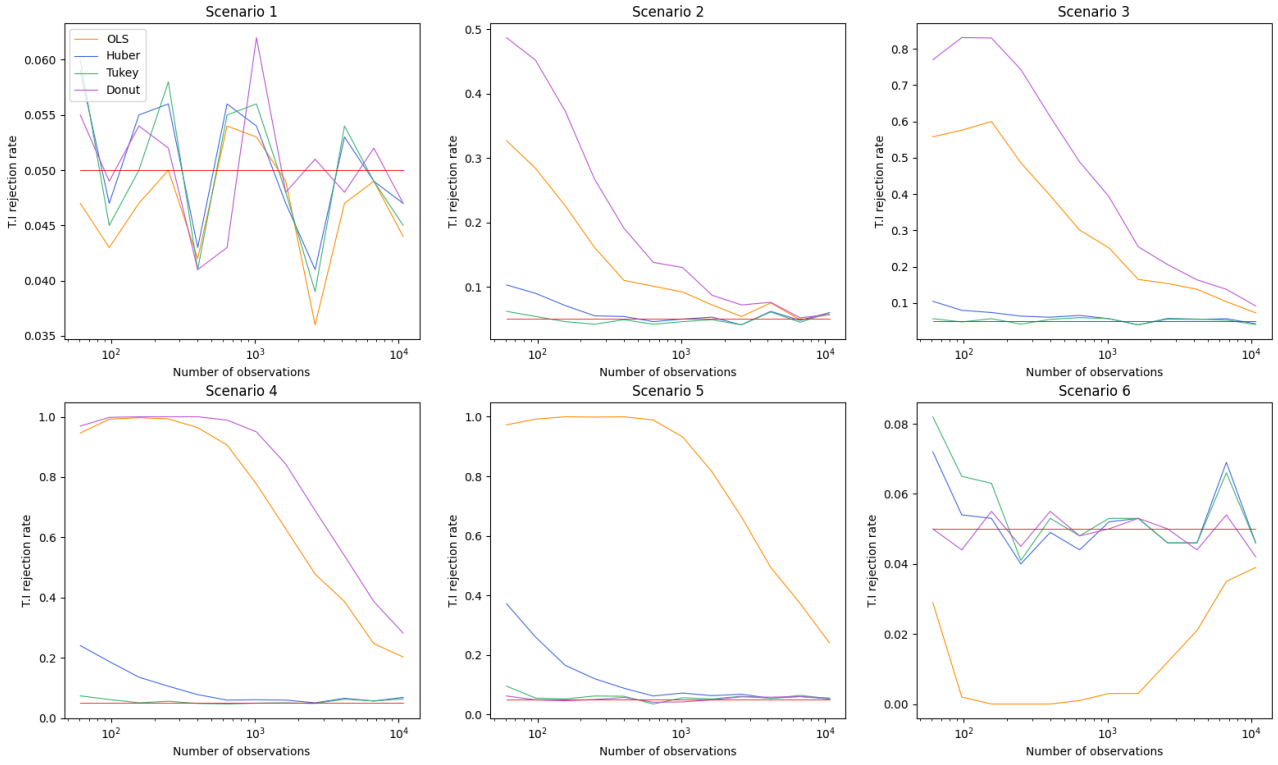

Figure A.8: T.I error ($H_0 : \hat{\tau} = \tau$) of the ATE estimates based on the 4 estimation methods, for increasing sample sizes. From the simulation with $r = 1,000$, $\tau = -0.5$.

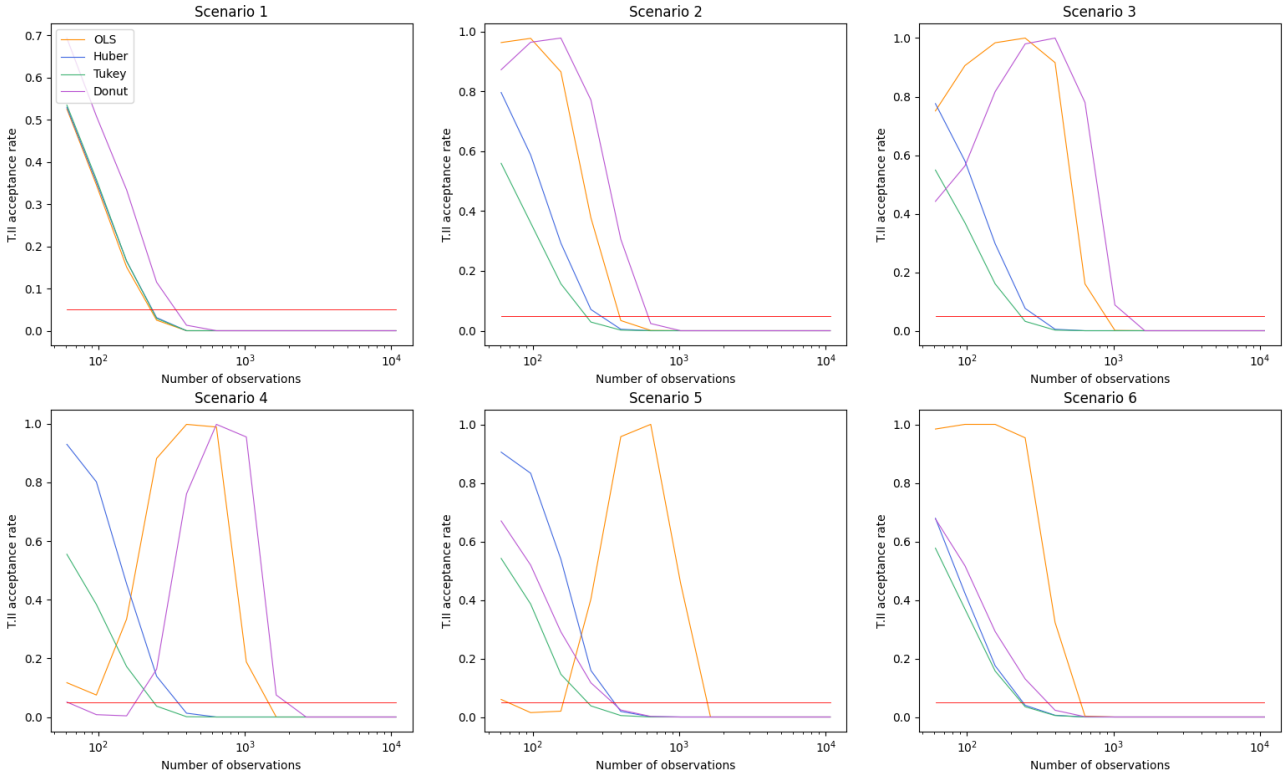

Figure A.9: T.II error ($H_0 : \hat{\tau} = 0$) of the ATE estimates from the 4 estimation methods, for increasing sample sizes. From the simulation with $r = 1,000$, $\tau = -0.5$.

A.2 Results from simulation with $\tau = 0$

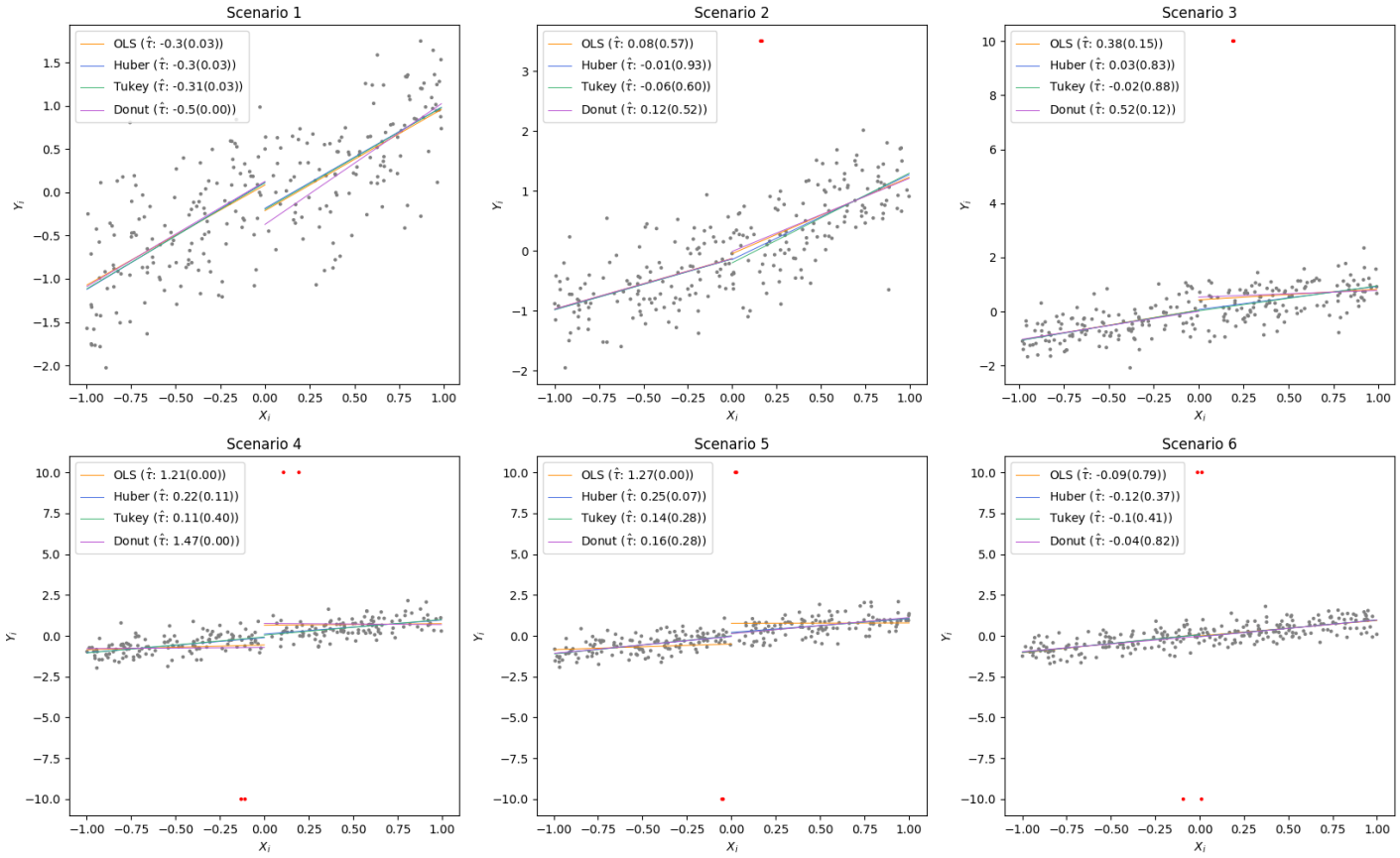

Figure A.10: The 6 different types of samples generated for each of the (outliers) scenarios with the fitted linear regressions estimated with the 4 different methods. The estimated ATE are also reported with their significance's. These are the first samples of each scenario from the simulation with $\tau = 0$.

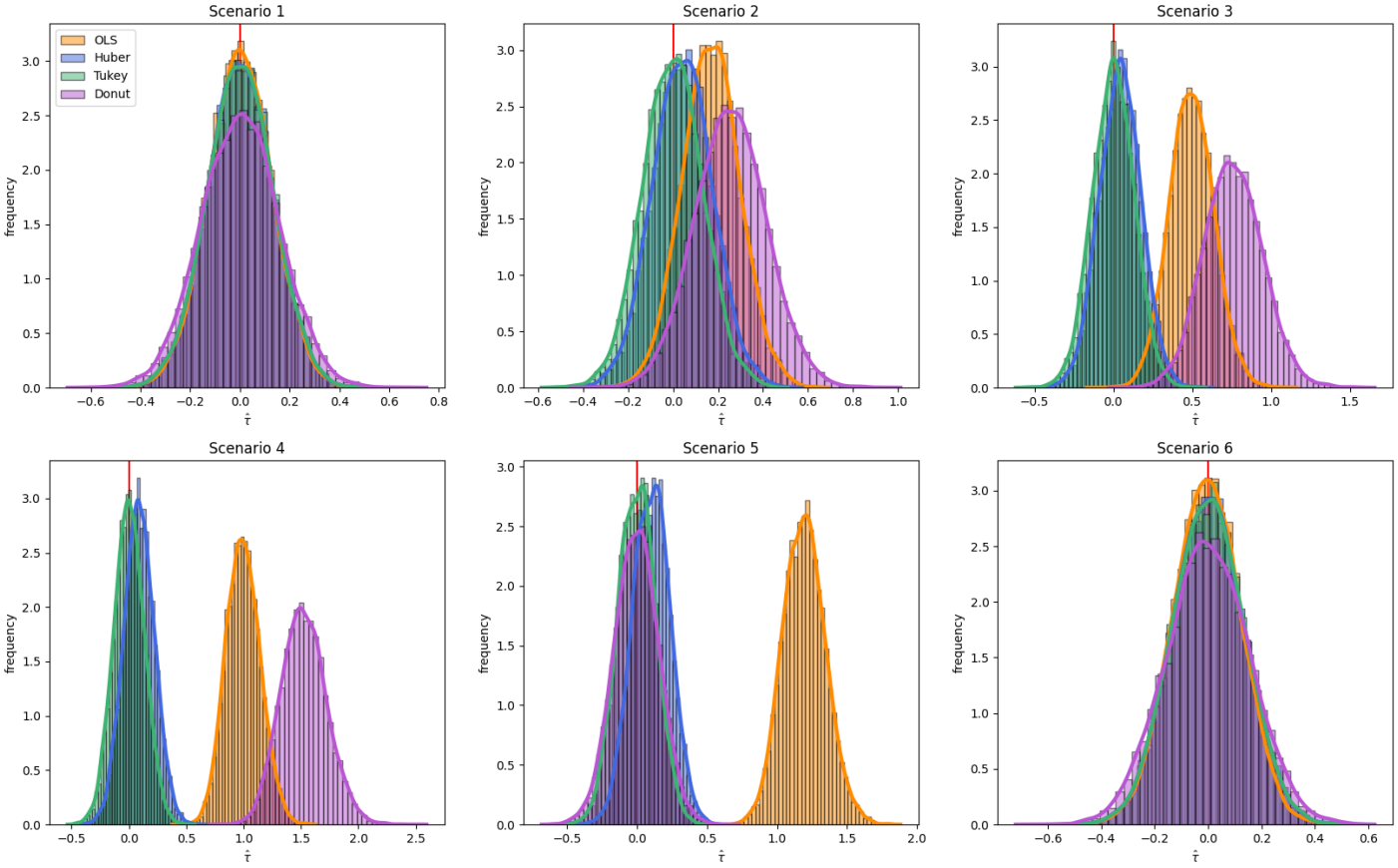

Figure A.11: Histograms of the estimated ATE based on the 4 different RDD estimation methods, for each of the (outliers) scenarios. From the simulation with $r = 10,000$, $\tau = 0$.

Table A.5: Bias, standard deviation and root mean squared error of the point estimates of the treatment effect. From the simulation with $r = 10,000$, $\tau = 0$.

	Scenario 1			Scenario 2			Scenario 3		
	Bias	St.Dev.	RMSE	Bias	St.Dev.	RMSE	Bias	St.Dev.	RMSE
OLS	0.001	0.128	0.128	0.168	0.128	0.211	0.498	0.142	0.517
Huber	0.001	0.132	0.132	0.042	0.133	0.139	0.044	0.132	0.139
Tukey	0.001	0.132	0.132	-0.001	0.133	0.133	0.000	0.132	0.132
Donut	0.001	0.157	0.157	0.256	0.159	0.302	0.761	0.186	0.784
	Scenario 4			Scenario 5			Scenario 6		
	Bias	St.Dev.	RMSE	Bias	St.Dev.	RMSE	Bias	St.Dev.	RMSE
OLS	0.996	0.149	1.007	1.194	0.152	1.204	-0.005	0.126	0.127
Huber	0.088	0.134	0.160	0.102	0.135	0.169	0.001	0.134	0.134
Tukey	0.001	0.134	0.134	-0.003	0.134	0.134	0.001	0.134	0.134
Donut	1.525	0.199	1.538	-0.002	0.159	0.159	-0.000	0.156	0.156

Table A.6: Skewness, kurtosis and p-value of the jarque-bera statistic of the estimated treatment effects. From the simulation with $r = 10,000$, $\tau = 0$.

	Scenario 1			Scenario 2			Scenario 3		
	Skew	Kurt	JB	Skew	Kurt	JB	Skew	Kurt	JB
OLS	0.008	2.993	0.942	0.011	3.025	0.799	0.082	3.107	0.000
Huber	-0.003	2.997	0.991	-0.002	3.016	0.944	-0.023	3.040	0.465
Tukey	-0.001	2.999	0.999	-0.006	3.002	0.974	-0.034	3.060	0.180
Donut	0.012	3.043	0.606	0.047	3.071	0.057	0.125	3.138	0.000
	Scenario 4			Scenario 5			Scenario 6		
	Skew	Kurt	JB	Skew	Kurt	JB	Skew	Kurt	JB
OLS	0.136	2.972	0.000	0.139	3.030	0.000	-0.019	3.028	0.622
Huber	0.026	2.991	0.562	-0.028	2.903	0.073	0.011	2.988	0.878
Tukey	0.012	2.988	0.856	-0.042	2.910	0.043	0.011	2.985	0.866
Donut	0.193	3.117	0.000	-0.009	3.053	0.517	0.008	2.984	0.902

Table A.7: Correct coverage of the confidence intervals for τ and their length. For significance level of $\alpha = 0.05$. From the simulation with $r = 10,000$, $\tau = 0$.

	Scenario 1		Scenario 2		Scenario 3	
	C.C.	Length	C.C.	Length	C.C.	Length
OLS	0.947	0.502	0.833	0.582	0.518	1.007
Huber	0.945	0.512	0.938	0.521	0.939	0.519
Tukey	0.946	0.513	0.947	0.515	0.949	0.514
Donut	0.952	0.620	0.751	0.727	0.264	1.286
	Scenario 4		Scenario 5		Scenario 6	
	C.C.	Length	C.C.	Length	C.C.	Length
OLS	0.006	1.333	0.000	1.334	1.000	1.374
Huber	0.899	0.528	0.881	0.526	0.952	0.526
Tukey	0.947	0.517	0.945	0.515	0.946	0.516
Donut	0.000	1.711	0.949	0.620	0.953	0.620

Table A.8: Type I and type II errors of t-test for $h_0 : \hat{\tau} = \tau$ and $h_0 : \hat{\tau} = 0$ repectively, from the estimated treatment effects. From the simulation with $r = 10,000$, $\tau = 0$. For significance level of $\alpha = 0.05$. Note that in this case T.I $= 1-$ T.II .

	Scenario 1		Scenario 2		Scenario 3	
	T.I	T.II	T.I	T.II	T.I	T.II
OLS	0.053	0.947	0.167	0.833	0.481	0.518
Huber	0.055	0.945	0.062	0.938	0.061	0.939
Tukey	0.054	0.946	0.053	0.947	0.051	0.949
Donut	0.048	0.952	0.249	0.751	0.736	0.264
	Scenario 4		Scenario 5		Scenario 6	
	T.I	T.II	T.I	T.II	T.I	T.II
OLS	0.994	0.006	1.000	0.000	0.000	1.000
Huber	0.101	0.899	0.119	0.881	0.048	0.952
Tukey	0.053	0.947	0.055	0.945	0.054	0.946
Donut	1.000	0.000	0.051	0.949	0.047	0.953

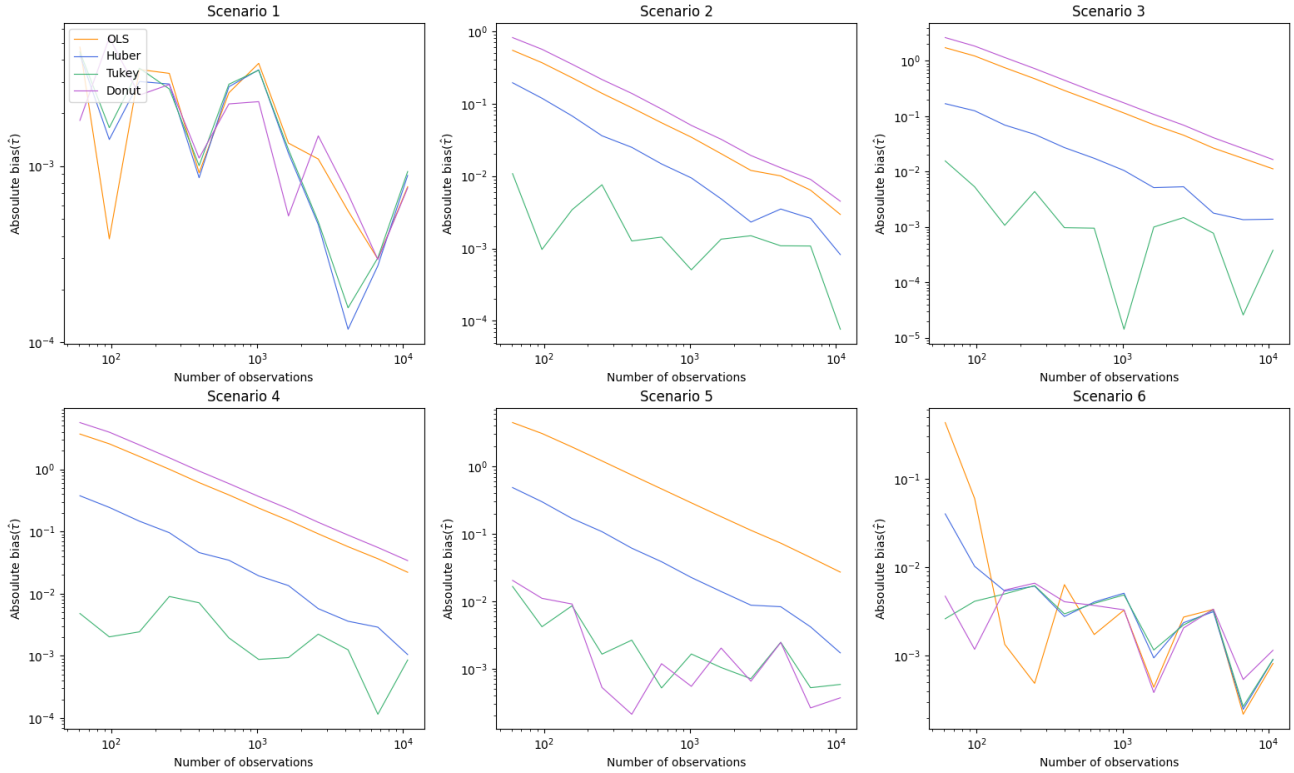

Figure A.12: Absolute value of the bias from the ATE estimates from the 4 estimation methods for increasing sample sizes. From the simulation with $r = 1,000$, $\tau = 0$.

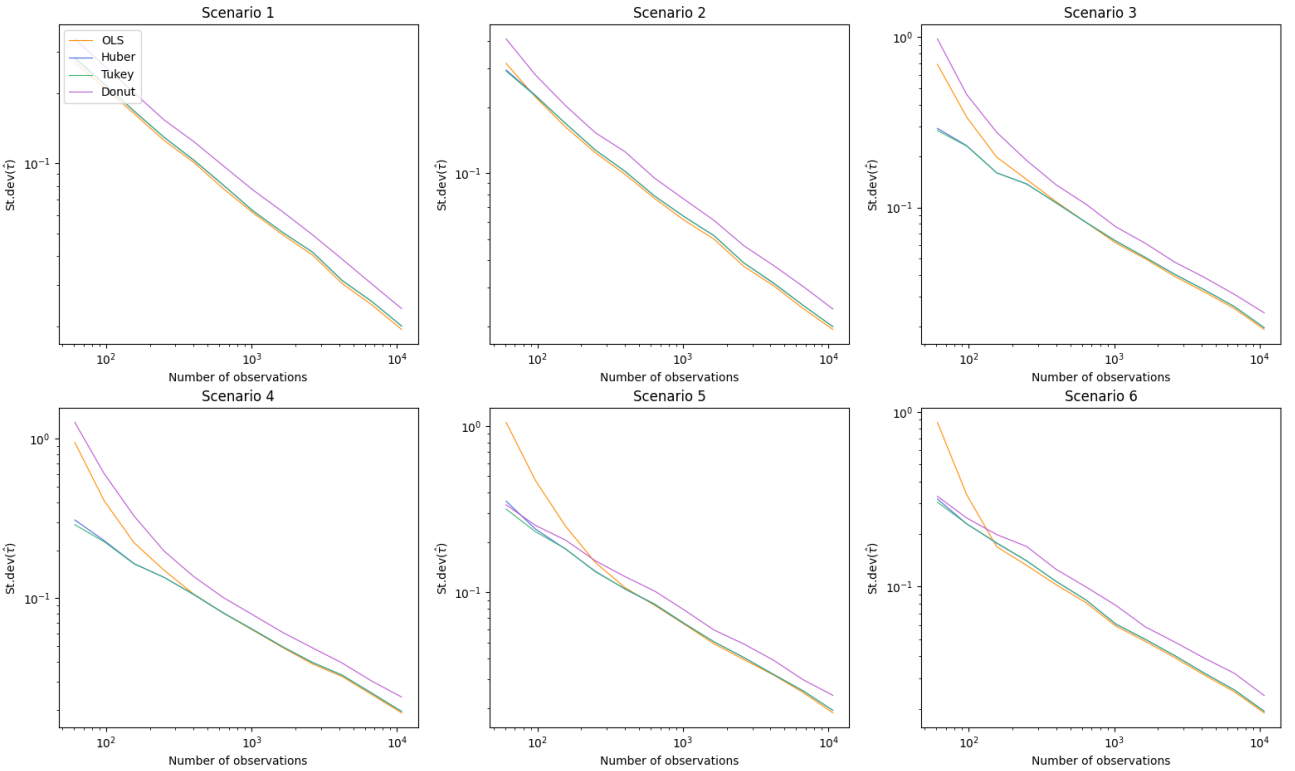

Figure A.13: Standard deviation of the ATE estimates from the 4 estimation methods for increasing sample sizes. From the simulation with $r = 1,000$, $\tau = 0$.

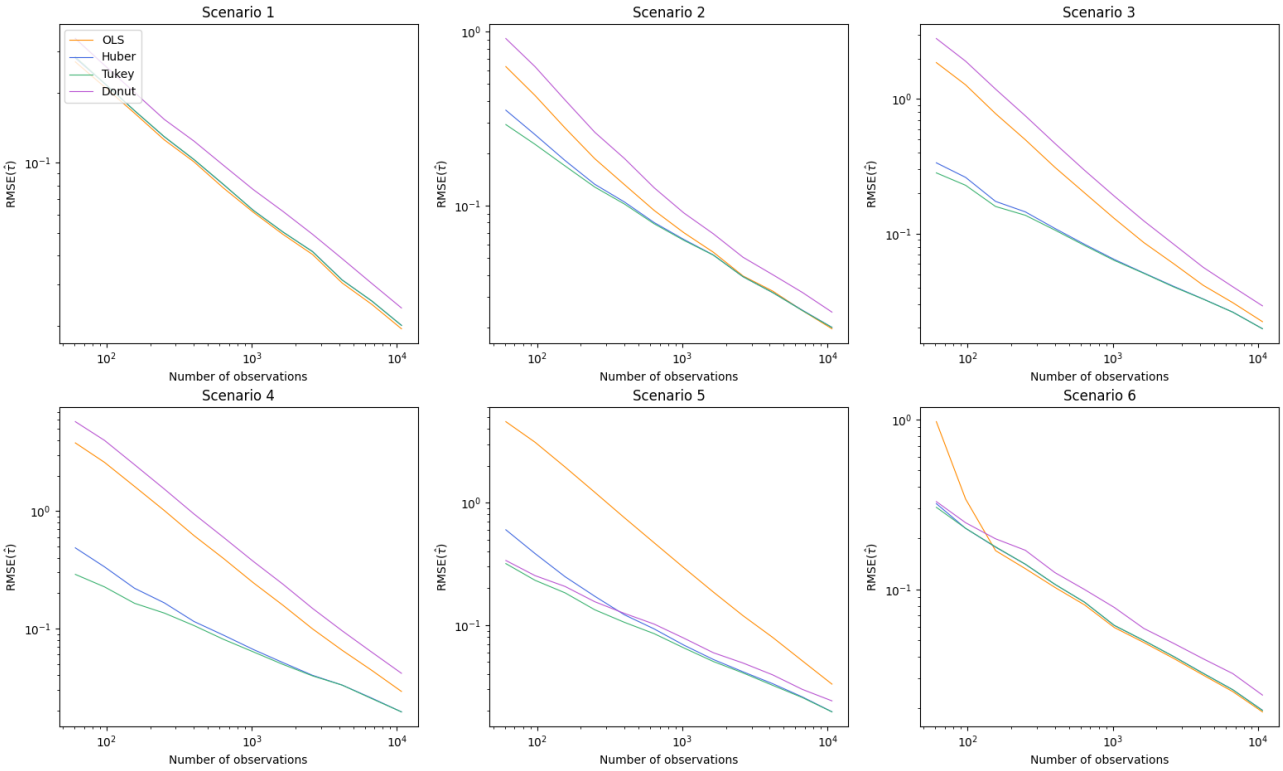

Figure A.14: RMSE of the ATE estimates from the 4 estimation methods for increasing sample sizes. From the simulation with $r = 1,000$, $\tau = 0$.

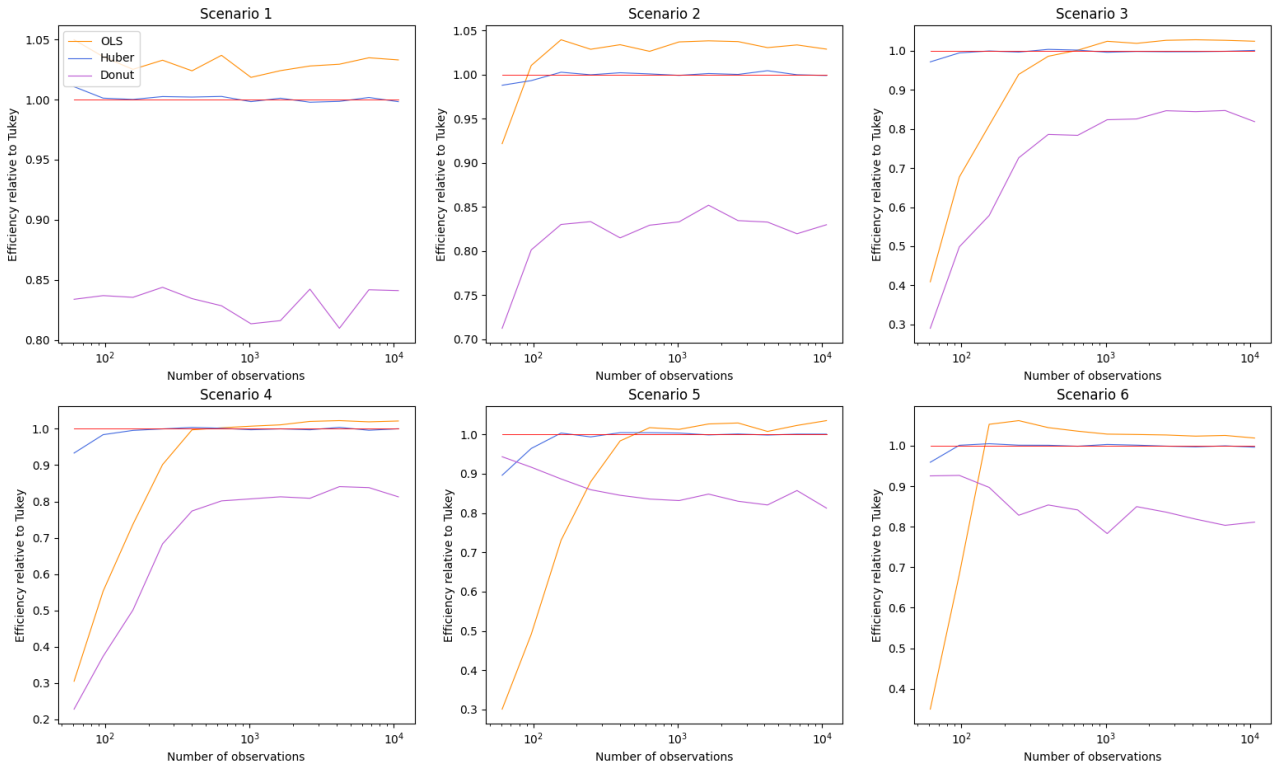

Figure A.15: Efficiency of the ATE estimates from the 3 estimation methods relative to Tukey estimates, for increasing sample sizes. Values above 1 mean that the estimate is more efficient than Tukey's. From the simulation with $r = 1,000$, $\tau = 0$.

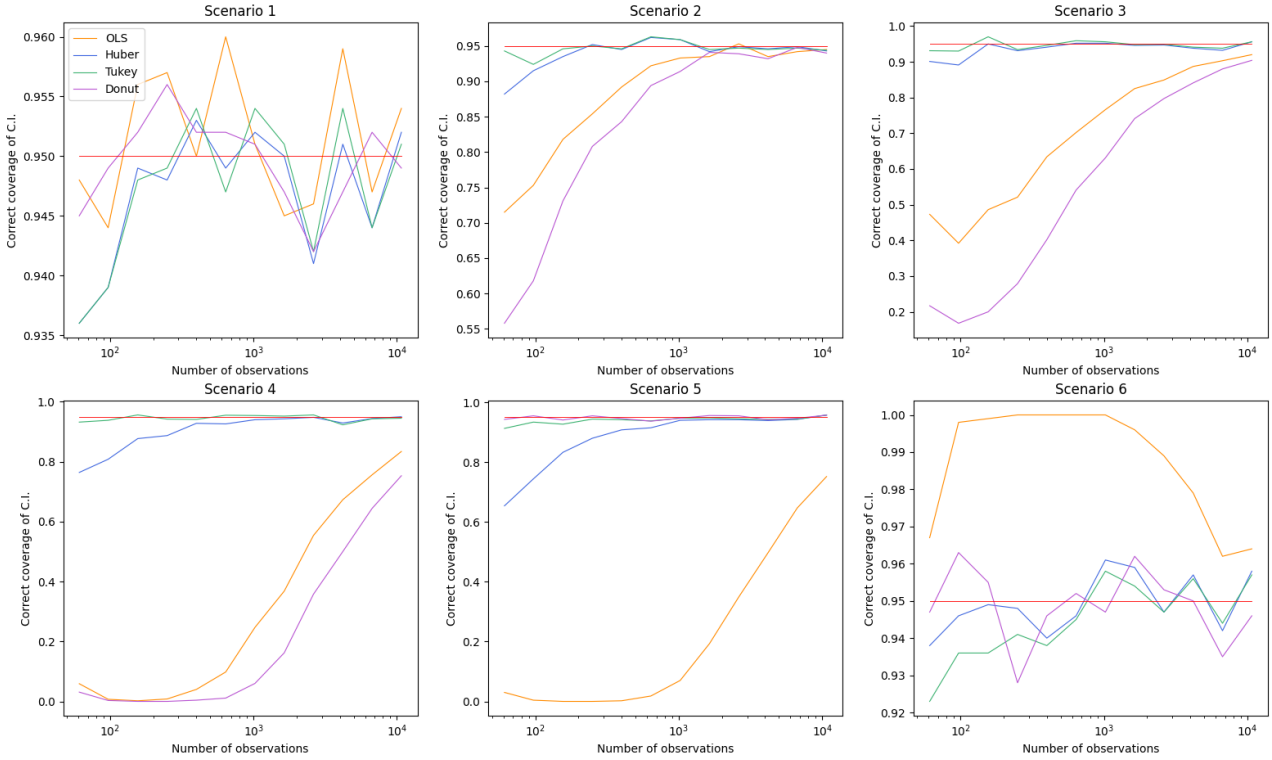

Figure A.16: Correct coverage of the confidence intervals of the ATE estimates from the 4 estimation methods, for increasing sample sizes. From the simulation with $r = 1,000$, $\tau = 0$.

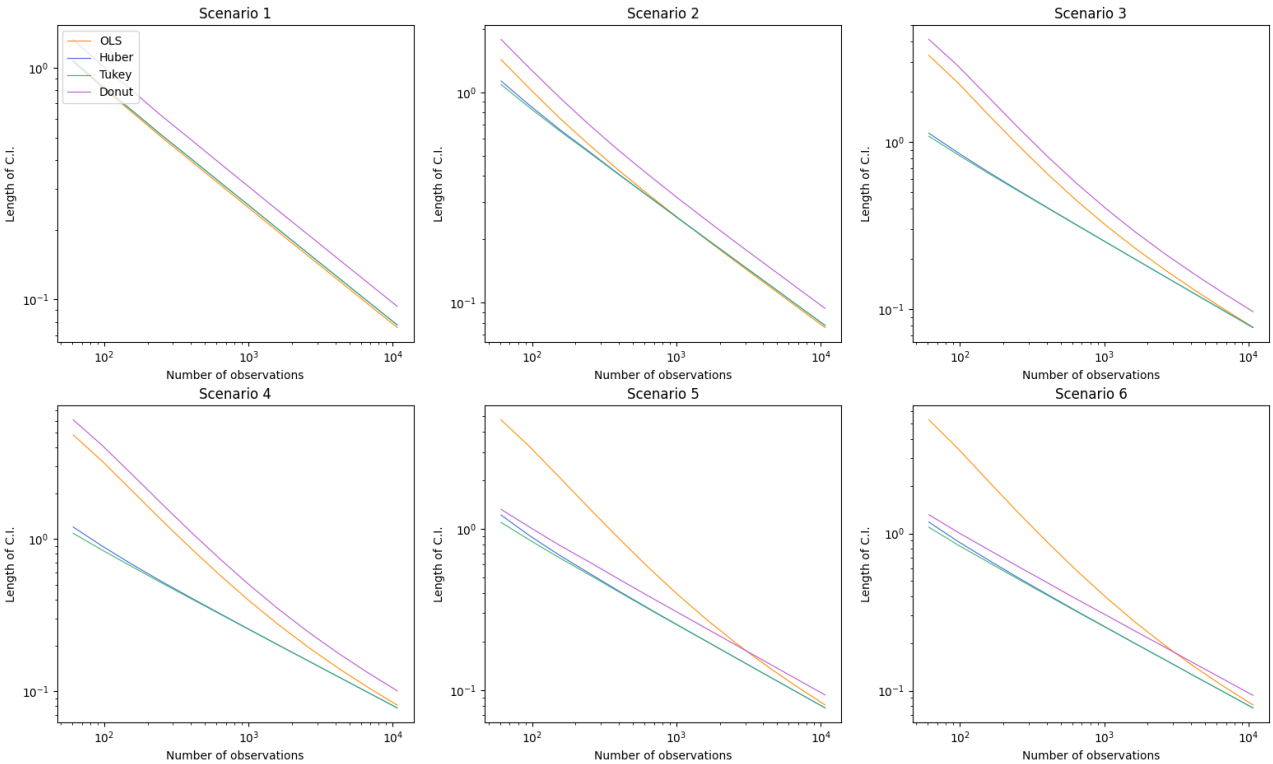

Figure A.17: Length of the confidence intervals of the ATE estimates from the 4 estimation methods, for increasing sample sizes. From the simulation with $r = 1,000$, $\tau = 0$.

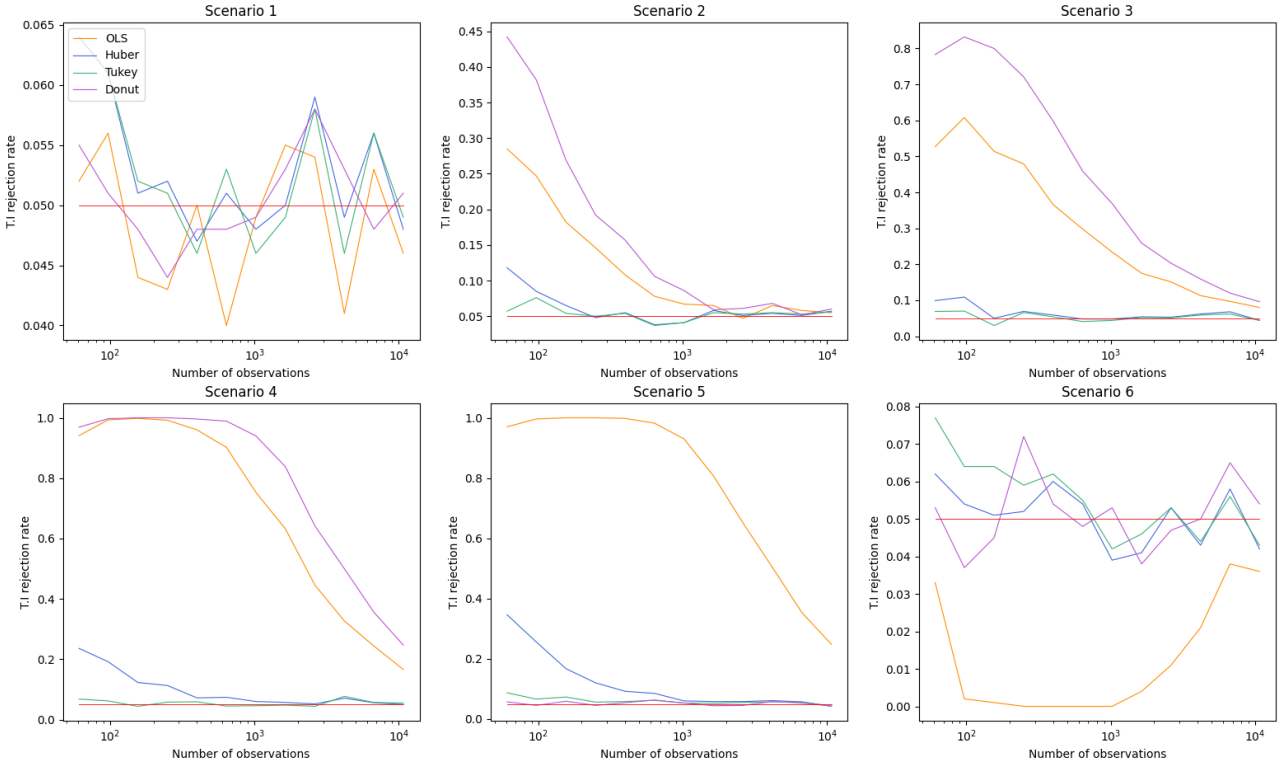

Figure A.18: T.I error ($H_0 : \hat{\tau} = \tau$) of the ATE estimates based on the 4 estimation methods, for increasing sample sizes. From the simulation with $r = 1,000$, $\tau = 0$.

B Replication

For the replication of the simulation from Noack and Rothe (2023), the same $\mu_L(X_i)$ is used with the GDP outlined in section 3.1:

$$\mu_{Noack,L}(X_i) = sign(X_i)X_i^2 - Lsign(X_i)\Big((X_i - \delta sign(X_i))^2 - \delta^2 sign(X_i)\Big)\mathbf{1}\{|X_i| < \delta\} \quad (7)$$

With $\delta = 0.1$ and $L = \{0, 10, 20, 30, 40, 200\}$,[8] the different levels of L are named by scenarios 1 to 6 , and $r = 1,000$. The methods developed for this thesis (Tukey and Huber RDD) were also used, and the bandwidth was fixed at 0.35.

Table B.1: Bias, standard deviation and root mean squared error of the point estimates of the treatment effect. From the replication simulation with $r = 1000$, $\tau = 0$.

	Scenario 1			Scenario 2			Scenario 3		
	Bias	St.Dev.	RMSE	Bias	St.Dev.	RMSE	Bias	St.Dev.	RMSE
OLS	-0.041	0.107	0.114	-0.106	0.109	0.152	-0.177	0.110	0.209
Huber	-0.040	0.110	0.117	-0.105	0.113	0.154	-0.178	0.114	0.211
Tukey	-0.040	0.110	0.117	-0.105	0.113	0.154	-0.178	0.114	0.212
Donut	-0.079	0.211	0.226	-0.083	0.212	0.228	-0.085	0.210	0.226
	Scenario 4			Scenario 5			Scenario 6		
	Bias	St.Dev.	RMSE	Bias	St.Dev.	RMSE	Bias	St.Dev.	RMSE
OLS	-0.250	0.113	0.275	-0.308	0.110	0.327	-1.391	0.200	1.405
Huber	-0.251	0.117	0.277	-0.310	0.113	0.330	-1.460	0.233	1.479
Tukey	-0.252	0.116	0.277	-0.310	0.114	0.330	-1.459	0.234	1.477
Donut	-0.097	0.216	0.236	-0.086	0.201	0.219	-0.100	0.211	0.234

Table B.2: Correct coverage of the confidence intervals for τ and their length. For significance level of $\alpha = 0.05$. From the replication simulation with $r = 1000$, $\tau = 0$.

	Scenario 1		Scenario 2		Scenario 3	
	C.C.	Length	C.C.	Length	C.C.	Length
OLS	0.934	0.424	0.825	0.424	0.613	0.426
Huber	0.931	0.434	0.837	0.433	0.627	0.436
Tukey	0.935	0.434	0.832	0.434	0.622	0.436
Donut	0.934	0.825	0.920	0.824	0.932	0.821
	Scenario 4		Scenario 5		Scenario 6	
	C.C.	Length	C.C.	Length	C.C.	Length
OLS	0.368	0.432	0.205	0.437	0.000	0.693
Huber	0.394	0.442	0.228	0.448	0.000	0.716
Tukey	0.396	0.442	0.230	0.448	0.001	0.715
Donut	0.920	0.825	0.943	0.824	0.908	0.818

[8] L=200 is reported as an extra scenario

C Programming Code

The Code for this thesis was developed in python 3.12.2 . A public repository containing the code base can be found at https://github.com/FranciscoPortilha/BSc2-Thesis—Robust-RDD, and a branch named Thesis-Final-Version will be kept unchanged with the code used for this thesis. The public libraries used are numpy (for handling arrays, and generating random variables), pandas (for managing dataframes), matplotlib (for making figures), statsmodels (for OLS and RLS regression models, some evaluation measures, and t-tests).

The code base contains 5 folders and 4 main files: application, final images, images, src, tests, mainSim.py, mainSimAdv.py, mainApplication.py and mainReplication.py. The application folder has the data for the application. In the final images folder the images used in the thesis are stored in 4 sub-folders (one for each main file). The images folder contains 4 empty sub-folders were the images from a run will be stored (re-running will overwrite these images). The methods developed for this thesis are in the src folder. And the tests folder contains some basic tests.

The file mainSim.py runs the base simulation, used for the simulations with $\tau = -0.5, 0$. mainSimAdv.py runs the power and the asymptotic simulations. mainApplication.py and mainRepplication.py are self explanatory. The file dataETL.py extracts the birth data from the raw data files, and also cluster it at gram level.

The src folder contains 5 files: sample.py, rrdd.py, simulation.py, simMetrics.py and exports.py. The sample.py file contains the methods for the generating the samples (with outliers) according to the DGP's for the main simulation and the replication study. The core method is genSample() which returns a sample randomly generated from desired DGP. The rrdd.py file has the methods to estimate RDD with the different estimation methods. simMetrics.py prepares the results to be exported, and exports.py contains the methods to create the figures and latex tables.

To ensure reproducability the main files are seeded a random number such that the result should be exactly the same when re-running the code, even if random.

List of Figures

List of Tables