Universität Leipzig
Fakultät für Mathematik und Informatik
Mathematisches Institut

Regression Discontinuity Design with Covariates

Diplomarbeit

Leipzig, 18. April 2023

vorgelegt von
Patrick Kramer
im Studiengang Mathematik Diplom

Betreuer: Jun.-Prof. Dr. Alexander Kreiß
Abteilung: Stochastik

I dedicate this thesis

. . . to my father who could not see this thesis completed. Thank you for always believing in me. I miss and love you.

. . . to my mother. Your endless support and encouragement enabled me the possibility to study. You are the reason for what I have achieved so far. Words can hardly express my gratitude and appreciation to you.

. . . to my sister. Thank you for being the best sister anyone can ever ask for. We are a team forever and ever.

Contents

Abstract

This thesis studies regression discontinuity designs with the use of additional covariates for estimation of the average treatment effect. We prove asymptotic normality of the covariate-adjusted estimator under sufficient regularity conditions. In the case of a high-dimensional setting with a large number of covariates depending on the number of observations, we discuss a Lasso-based selection approach as well as alternatives based on calculated correlation thresholds. We present simulation results on those alternative selection strategies.

1 Introduction

Regression discontinuity design (RDD) is regularly used in economics, political science and other social sciences like biomedical science. The aim of RDD is to estimate causal effects of a treatment on some outcome in settings where the treatment is applied depending on some running variable or score value. Therefore, given data needs to contain a score value for each unit which decides whether the treatment is applied to that unit or not. In fact, if the score value lies above a certain cutoff point, the treatment is applied, otherwise it is not. Then, the effect of the treatment on the outcome can be estimated by comparing units that are close to the cutoff, but are on different sides of it. This concept, including its theoretical properties and applications, has already been studied by a variety of researchers (e.g. in [HTdK01], [IL08], [LL10], [IK12], [CCT14], [CTVB17], [AK18], [CIT19] and [CIT23]).

In practice, there is often more data than only the score value and the outcome. For this reason, many researchers want to incorporate additional data into the analysis of the causal effect. Indeed, this can be done by including additional covariates linearly in a local linear regression around the cutoff as suggested by [CCFT19]. In this paper, Calonico et al. provide a formal study of the impact of covariate adjustments on estimation and inference in RD designs. In order to account for misspecification in finite samples [CCT14], they established a bias-corrected version of the covariate-adjusted estimator. In addition to that, smoothness assumptions that are necessary to obtain a consistent estimator that delivers reliable results are given. Under this imposed regularity conditions, they showed that the bias-corrected covariate-adjusted estimator is asymptotically normally distributed. However, what can also be seen is that the estimation just provides accurate results in low-dimensional settings and breaks down in case of a rather large number of covariates in proportion to the number of observations. Hence, scenarios in which rich data sets occur or a large number of transformations is applied on a small set of covariates are not covered by this theory.

Given a large amount of covariates that can even change with the number of observations, [KR21] addressed this problem by establishing a two-step approach. In the first step, a small subset of these covariates is selected by adding a Lasso penalty to the least squares problem defining the local linear adjustment estimator. The selection procedure

ensures that the covariates are strongly related to the outcome in order to obtain a smaller variance in the estimation. Then, in a second step, the selected covariates are included linearly in the local linear regression around the cutoff. Kreiß and Rothe showed that under an approximate sparsity condition, which informally means that only a small number of covariates is particularly relevant for the analysis, the post-Lasso estimator is asymptotically normally distributed with asymptotic bias and variance being similar as in the case of a low-dimensional setting.

In this thesis, we also want to consider RDD using additional covariates. The main goal is to prove that the covariate-adjusted estimator using a fixed number of covariates is asymptotically normally distributed under some necessary regularity condition being similar to those imposed in [CCFT19]. However, the main difference to [CCFT19] will be that we show the asymptotic normality for the non-bias-corrected version of the estimator. Our approach for that will be to adjust the proof of [KR21] to the case of a fixed number of covariates. Furthermore, the second goal of this thesis will tackle the situation of having potentially many covariates, i.e. when the number of covariates can depend on the number of observations. In fact, the goal is to make suggestions for alternatives to the covariate selection via including a Lasso penalty into the least squares problem. Our methods will rely on selecting covariates according to their correlation to the outcome. We will do this by calculating a threshold for each of the covariates. When the correlation coefficient of the respective covariate lies above the threshold, we will take it into consideration, else we will reject it. Also, we will tackle the situation when some of the chosen covariates are correlated to each other by establishing two different deletion procedures. After the selection, all chosen covariates will then be included linearly into the local linear regression similarly as after the Lasso-based selection.

This thesis contributes to the wide area of literature considering regression discontinuity designs using covariates, that, besides [CCFT19] and [KR21], also includes [AK18], [FH19], [NOR21] and [AOS21].

The remaining part of this thesis is structured as follows: Section 2 covers all necessary preliminaries in which we want to recall notation as well as theorems that are relevant in later chapters. In Section 3 we introduce the concept of regression discontinuity designs. We explain the framework, different types of RDD as well as how local regression is applied to estimate the average treatment effect. In the course of this, we will also comment on MSE-optimal bandwidth choices. We end the section by giving an outlook on incorporating covariates. With Section 4 the main part of the thesis begins as it sets out all the necessary terminology and notation as well as declares the fundamental RDD environment which is required for all subsequent statements. As the proof of the asymptotic normality of our covariate-adjusted estimator concludes in two steps, each one of them is covered within its own section. Section 5 covers the first step, namely it provides us with statements leading to a representation of the bias. Then, Section 6 covers the second major step as its aim is to show the asymptotic normality of the covariate-adjusted estimator. Section 7 first imposes all assumptions in one spot which are necessary so that the asymptotic normality

holds. After that, the main theorem of the thesis, which exactly states that asymptotic normality, is shown. The section ends by discussing the asymptotic bias and variance of the estimator and comparing them to those obtained in the case without using covariates. Chapter 8 proceeds by introducing the situation of having potentially many covariates, i.e. when the number of covariates can change with the number of observations. It describes how the covariate selection procedure implementing a Lasso penalty solves this. Finally, suggestions for alternatives to the Lasso-based selection are given. Section 9 contains the results of our conducted simulations. Section 10 draws a conclusion for future work, which is followed by a list of references.

2 Preliminaries

This chapter defines basic terminology that is frequently used throughout the thesis. Also, we want to provide Jensen's inequality as well as Lyapunov's Central Limit Theorem since we will need those in some of the subsequent proofs.

As our notation will involve a lot of indices, we will have the case that we want to refer to a component of a vector, let us say v_i, that already has an index i. To avoid misunderstandings, we will denote the k-th component of v_i by $v_i^{(k)}$ in those cases.

Moreover, since we will deal with asymptomatic behaviour and estimation, the Bachmann-Landau notation will occur quiet often. For this reason, we will give a brief overview of that concept in the following, whereby we will introduce the notation directly for sequences and leave out the more general case of functions.

Big-O Notation Let $(x_n)_{n\in\mathbb{N}}$ be a real (or complex) valued sequence and $(a_n)_{n\in\mathbb{N}}$ be a real valued comparison sequence. To describe that the behaviour of x_n is asymptotically bounded by a_n, we write

$$x_n = O(a_n) \text{ as } n \to \infty$$

if there exist $N \in \mathbb{N}$ and a constant $C > 0$ such that

$$|x_n| \leq Ca_n \text{ for all } n \geq N.$$

Little-o Notation An even stronger statement would be that x_n is asymptotically dominated by a_n. We can formulate this by writing

$$x_n = o(a_n) \text{ as } n \to \infty$$

if for all $\epsilon > 0$ there is $N \in \mathbb{N}$ such that we have for all $n \geq N$

$$|x_n| < \epsilon a_n.$$

Later on, we will use the notations $O_P(a_n)$ and $o_P(a_n)$ whereby we find the index P in addition. This means that we examine the asymptotic behaviour of random variables in the manner of convergence in probability. Indeed, this is what will be defined next.

Big-O in Probability Notation Given a sequence $(X_n)_{n\in\mathbb{N}}$ of random variables and a corresponding sequence of constants $(a_n)_{n\in\mathbb{N}}$, we write

$$X_n = O_P(a_n) \text{ as } n \to \infty$$

if for every $\epsilon > 0$, there exist $M > 0$ and $N \in \mathbb{N}$ such that for all $n \geq N$

$$P\left(\left|\frac{X_n}{a_n}\right| > M\right) < \epsilon.$$

Little-o in Probability Notation Given the sequences $(X_n), (a_n)$ as above, we write

$$X_n = o_P(a_n) \text{ as } n \to \infty$$

if for all $\epsilon > 0$

$$P\left(\left|\frac{X_n}{a_n}\right| \geq \epsilon\right) \xrightarrow[n\to\infty]{} 0.$$

Sometimes, we also want to consider $O(a_n), o(a_n), O_P(a_n), o_P(a_n)$ as sets of sequences that satisfy the respective asymptotic condition. This allows us to write for instance $o(a_n) \subseteq O(a_n)$ in order to state that every sequence of order $o(a_n)$ is also of order $O(a_n)$. Moreover, given a sequence of vectors or matrices M_n, we also want to use Bachmann-Landau notation for that case. For example, if we write $M_n = O_P(1)$, we just interpret the absolute value in the definition of O_P as Euclidean norm or its induced matrix norm, respectively. Yet, we could also use any other norm since they are equivalent in the finite-dimensional case.

We assume the basic properties of the Bachmann-Landau notation as known. However, we want to highlight some essential properties of the notation in probability as this will be of great importance for subsequent proofs.

Lemma 2.1 (Properties of Bachmann-Landau in Probability)**.** *Let a_n and b_n be sequences of constants. The following statements hold:*

(a) $O(a_n) \subseteq O_P(a_n)$ *and* $o(a_n) \subseteq o_P(a_n)$

(b) If there exists $N \in \mathbb{N}$ such that $b_n \geq a_n$ for all $n \geq N$, then $O_P(a_n) \subseteq O_P(b_n)$ as well as $o_P(a_n) \subseteq o_P(b_n)$

(c) $O_P(a_n) + O_P(a_n) = O_P(a_n)$ *and* $o_P(a_n) + o_P(a_n) = o_P(a_n)$

(d) $O_P(a_n)O_P(b_n) = O_P(a_n b_n)$ *and* $o_P(a_n)o_P(b_n) = o_P(a_n b_n)$

(e) $o_P(a_n)O_P(b_n) = o_P(a_n b_n)$

(f) If $a_n \to 0$, then $O_P(a_n) = o_P(1)$

(g) Let $A_n \in \mathbb{R}^{p\times p}$ be invertible for all $n \in \mathbb{N}$, (a_n) be a sequence converging to zero and $A_n^{-1} = O(1), X_n = O_P(a_n)$ with $(A_n + X_n)^{-1} = O_P(1)$, then $(A_n + X_n)^{-1} = A_n^{-1} + O_P(a_n)$.

(h) Let $A_n \in \mathbb{R}^{p\times p}$ be invertible for all $n \in \mathbb{N}$ and $A_n^{-1} = O(1), X_n = o_P(1)$ with $(A_n + X_n)^{-1} = O_P(1)$, then $(A_n + X_n)^{-1} = A_n^{-1} + o_P(1)$.

Proof. (a) Let $x_n = O(a_n)$. Then, by definition, there exist $N \in \mathbb{N}$ and $C > 0$ such that $|x_n| \leq C a_n$ for all $n \geq N$. For $\epsilon > 0$ choose $M \geq C$. Then for all $n \geq N$

$$P\left(\left|\frac{x_n}{a_n}\right| > M\right) \leq P\left(\left|\frac{x_n}{a_n}\right| > C\right) \leq P\left(C > C\right) = 0 < \epsilon$$

leading to $x_n = O_P(a_n)$.

Now, let $y_n = o(a_n)$. Then $\left|\frac{y_n}{a_n}\right| \to 0$, hence for arbitrary $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that $\left|\frac{y_n}{a_n}\right| < \epsilon$ for all $n \geq N$. This leads to

$$P\left(\left|\frac{y_n}{a_n}\right| \geq \epsilon\right) = 0$$

for all $n \geq N$, which means that $y_n = o_P(a_n)$.

(b) Let $N \in \mathbb{N}$ such that $a_n \leq b_n$ for all $n \geq N$. Then for all $M > 0$ and $n \geq N$

$$P\left(\left|\frac{X_n}{b_n}\right| > M\right) \leq P\left(\left|\frac{X_n}{a_n}\right| > M\right),$$

which proves both statements.

(c) Let $X_n = O_P(a_n)$ and $Y_n = O_P(a_n)$. For arbitrary $\epsilon > 0$ there exist $M_1, M_2 > 0$ and $N_1, N_2 \in \mathbb{N}$ such that

$$P\left(\left|\frac{X_n}{a_n}\right| > M_1\right) < \frac{\epsilon}{2} \quad \text{for all } n \geq N_1$$

and

$$P\left(\left|\frac{Y_n}{a_n}\right| > M_2\right) < \frac{\epsilon}{2} \quad \text{for all } n \geq N_2.$$

Set $M := 2\max(M_1, M_2)$, $N := \max(N_1, N_2)$. Then, for all $n \geq N$

$$\begin{aligned}
&P\left(\left|\frac{X_n + Y_n}{a_n}\right| > M\right) \\
&\leq P\left(\left|\frac{X_n}{a_n}\right| + \left|\frac{Y_n}{a_n}\right| > M\right) \\
&\leq P\left(2\left|\frac{X_n}{a_n}\right| > M \ \vee \ 2\left|\frac{Y_n}{a_n}\right| > M\right) \\
&\leq P\left(2\left|\frac{X_n}{a_n}\right| > M\right) + P\left(2\left|\frac{Y_n}{a_n}\right| > M\right) \\
&\leq P\left(\left|\frac{X_n}{a_n}\right| > M_1\right) + P\left(\left|\frac{Y_n}{a_n}\right| > M_2\right) \\
&< \frac{\epsilon}{2} + \frac{\epsilon}{2} \\
&= \epsilon
\end{aligned}$$

leading to $X_n + Y_n = O_P(a_n)$.

Now, let $X_n = o_P(a_n)$ and $Y_n = o_P(a_n)$. Then, analogously to the above inequalities,

we obtain for arbitrary $\epsilon > 0$ that

$$\begin{aligned} & P\left(\left|\frac{X_n + Y_n}{a_n}\right| \geq \epsilon\right) \\ \leq\ & P\left(\left|\frac{X_n}{a_n}\right| \geq \frac{\epsilon}{2}\right) + P\left(\left|\frac{Y_n}{a_n}\right| \geq \frac{\epsilon}{2}\right) \xrightarrow[n\to\infty]{} 0 \end{aligned}$$

as both summands converge to zero by definition of $o_P(a_n)$. Hence, $X_n + Y_n = o_P(a_n)$.

(d) Let $X_n = O_P(a_n)$ and $Y_n = O_P(b_n)$. For arbitrary $\epsilon > 0$ there exists $M_1, M_2 > 0$ and $N_1, N_2 \in \mathbb{N}$ such that

$$P\left(\left|\frac{X_n}{a_n}\right| > M_1\right) < \frac{\epsilon}{2} \quad \text{for all } n \geq N_1$$

and

$$P\left(\left|\frac{Y_n}{b_n}\right| > M_2\right) < \frac{\epsilon}{2} \quad \text{for all } n \geq N_2.$$

Set $M := \max(M_1, M_2)^2, N = \max(N_1, N_2)$. Then, for all $n \geq N$

$$\begin{aligned} & P\left(\left|\frac{X_n Y_n}{a_n b_n}\right| > M\right) \\ \leq\ & P\left(\left|\frac{X_n}{a_n}\right|^2 > M \ \vee\ \left|\frac{Y_n}{a_n}\right|^2 > M\right) \\ \leq\ & P\left(\left|\frac{X_n}{a_n}\right|^2 > M\right) + P\left(\left|\frac{Y_n}{a_n}\right|^2 > M\right) \\ \leq\ & P\left(\left|\frac{X_n}{a_n}\right| > M_1\right) + P\left(\left|\frac{Y_n}{a_n}\right| > M_2\right) \\ <\ & \frac{\epsilon}{2} + \frac{\epsilon}{2} \\ =\ & \epsilon, \end{aligned}$$

which results in $X_n Y_n = O_P(a_n b_n)$. Similarly, we would show that $o_P(a_n) o_P(b_n) = o_P(a_n b_n)$. As we do not need this particular statement in the thesis, we omit the proof here.

(e) Let $X_n = o_P(a_n)$ and $Y_n = O_P(b_n)$. Let $\tilde{\epsilon} > 0$, then we have to show that

$$\forall \epsilon > 0 \exists N \in \mathbb{N} \forall n \geq N : P\left(\left|\frac{X_n Y_n}{a_n b_n}\right| \geq \tilde{\epsilon}\right) < \epsilon. \tag{2.1}$$

So, let $\epsilon > 0$. By definition of $O_P(b_n)$, we know that there exist $M > 0$ and $\tilde{N} \in \mathbb{N}$ such that

$$P\left(\left|\frac{Y_n}{b_n}\right| > M\right) < \frac{\epsilon}{2}.$$

for all $n \geq \tilde{N}$. Now, by definition of $o_P(a_n)$, we know that there is $\hat{N} \in \mathbb{N}$ such that

$$P\left(\left|\frac{X_n}{a_n}\right| \geq \frac{\tilde{\epsilon}}{M}\right) < \frac{\epsilon}{2}.$$

for all $n \geq \hat{N}$. Set $N := \max(\tilde{N}, \hat{N})$. Then, for all $n \geq N$

$$\begin{aligned}
&P\left(\left|\frac{X_n Y_n}{a_n b_n}\right| \geq \tilde{\epsilon}\right) \\
&\leq P\left(\left|\frac{X_n}{a_n}\right| \geq \frac{\tilde{\epsilon}}{M} \ \vee \ \left|\frac{Y_n}{b_n}\right| > M\right) \\
&\leq P\left(\left|\frac{X_n}{a_n}\right| \geq \frac{\tilde{\epsilon}}{M}\right) + P\left(\left|\frac{Y_n}{b_n}\right| > M\right) \\
&< \frac{\epsilon}{2} + \frac{\epsilon}{2} \\
&= \epsilon.
\end{aligned}$$

Hence, we verified (2.1), which means that for all $\tilde{\epsilon} > 0$

$$P\left(\left|\frac{X_n Y_n}{a_n b_n}\right| \geq \tilde{\epsilon}\right) \xrightarrow[n\to\infty]{} 0.$$

This proves $X_n Y_n = o_P(a_n b_n)$.

(f) Let $X_n = O_P(a_n)$ whereby a_n converges to zero. Let $\tilde{\epsilon} > 0$, then we want to show that

$$\forall \epsilon > 0 \exists N \in \mathbb{N} \forall n \geq N : P\left(|X_n| > \tilde{\epsilon}\right) < \epsilon. \tag{2.2}$$

Therefore, let $\epsilon > 0$. By definition of $O_P(a_n)$, there exist $M > 0$ and $\tilde{N} \in \mathbb{N}$ such that

$$P\left(\left|\frac{X_n}{a_n}\right| > M\right) < \epsilon$$

for all $n \geq \tilde{N}$. Furthermore, choose $\hat{N} \in \mathbb{N}$ such that

$$M|a_n| < \tilde{\epsilon}$$

for all $n \geq \hat{N}$. This is possible since a_n converges to zero. Now set $N := \max(\tilde{N}, \hat{N})$, then for all $n \geq N$

$$P\left(|X_n| > \tilde{\epsilon}\right) \leq P\left(|X_n| > M|a_n|\right) < \epsilon,$$

which shows (2.2). This leads to $X_n = o_P(1)$.

(g) Let $A_n \in \mathbb{R}^{p\times p}$ be invertible, (a_n) be a sequence converging to zero and $A_n^{-1} = O(1)$,

$X_n = O_P(a_n)$ such that $(A_n + X_n)^{-1} = O(1)$. Then we have

$$\begin{aligned} A_n^{-1} - (A_n + X_n)^{-1} &= A_n^{-1}(I - A_n(A_n + X_n)^{-1}) = A_n^{-1}(A_n + X_n - A_n)(A_n + X_n)^{-1} \\ &= A_n^{-1}X_n(A_n + X_n)^{-1} = O(1)O_P(a_n)O_P(1) \\ &\overset{(a)}{=} O_P(1)O_P(a_n)O_P(1) \\ &\overset{(d)}{=} O_P(a_n), \end{aligned}$$

leading to the assertion being shown.

(h) We just refer to the proof of (g) since it is shown in an analogous manner. □

Now, we want to state Jensen's inequality. As we provide the assertion without proof, we refer for example to [Dur19] on how to show this.

Proposition 1 (Jensen's inequality)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, X an integrable real-valued random variable and φ a convex function. Then*

$$\varphi(\mathbb{E}(X)) \leq \mathbb{E}(\varphi(X))$$

whereby equality holds if and only if there is a convex set A such that $P(X \in A) = 1$ and φ is linear on A.

At last, we state Lyapunov's Central Limit Theorem, which is a variant of the classic Central Limit Theorem. Again, we omit the proof and refer to [Kle14].

Proposition 2 (Lyapunov's Central Limit Theorem)**.** *For every $n \in \mathbb{N}$ let $k_n \in \mathbb{N}$ and*

$$X_{n,1}, X_{n,2}, ..., X_{n,k_n}$$

be random variables such that $(X_{n,i})_{i=1,...,k_n}$ is an independent family of random variables for every $n \in \mathbb{N}$. Suppose that the mean $\mu_{n,i}$ and variance $\sigma_{n,i}^2$ of each $X_{n,i}$ is finite. Further, define

$$\mathcal{S}_n^2 = \sum_{i=1}^{k_n} \sigma_{n,i}^2$$

and assume that for some $\delta > 0$ Lyapunov's condition

$$\lim_{n\to\infty} \frac{1}{\mathcal{S}_n^{2+\delta}} \sum_{i=1}^{k_n} \mathbb{E}\left(|X_{n,i} - \mu_{n,i}|^{2+\delta}\right) = 0$$

holds. Then the sum with the summands $\frac{X_{n,i}-\mu_{n,i}}{\mathcal{S}_n}$ converges in distribution to the standard normal distribution, i.e.

$$\frac{1}{\mathcal{S}_n} \sum_{i=1}^{k_n} (X_{n,i} - \mu_{n,i}) \xrightarrow{d} \mathcal{N}(0, 1).$$

Next chapter will cover an introduction on regression discontinuity designs.

3 Regression Discontinuity Designs

The goal of this chapter is to provide a brief overview of regression discontinuity designs including the underlying statistical frameworks and methods for estimation and inference. For the sake of this introduction, we mainly follow the content of the presentation of Matias Cattaneo and Rocio Titiunik on the foundations of regression discontinuity designs, which was held at a conference of National Bureau of Economic Research [CT21]. Some additional information is also taken from the according paper [CIT19]. If the content is taken from other references, it is given in the respective paragraphs.

3.1 Designs and Frameworks

Regression discontinuity design (short: RDD) is a statistical design built to determine the effect of a so-called treatment variable on some outcome variable, whereby the treatment is applied depending on an observation (also called score) being above or below a certain threshold, the so-called cutoff. This means that entities having that observational variable greater than the threshold will receive the treatment, while the others will not.

This setting can be used to create a sample for which the treatment can be assumed to be applied randomly. In fact, this can be done by looking at a small interval around the cutoff as the entities that fall inside can be assumed to have similar characteristics since they are all close to the cutoff. In this area just a very small change in the score can cause the treatment to be applied or not. Therefore, we can assume it to be randomly applied around the cutoff point. Therefore, RDD is highly beneficial when the treatment cannot immediately be assumed to be received on a random basis.

Once we chose that interval around the cutoff, we can estimate the function of the outcome variable at the left and right side of the cutoff, respectively. This can be done by either local polynomial methods or so-called local randomization methods. Local randomization methods rely on assumptions regarding local random assignment of the treatment near the cutoff and use arguments from the study on analysis of experiments. However, throughout this introduction, we want to focus on local polynomial methods relying on continuity and differentiability assumptions. In fact, these methods perform local polynomial regression separately for the left and the right side of the cutoff. Then, the distance (difference in y-coordinates) between the endpoints of the regression functions at the cutoff can be measured and interpreted as the effect of the treatment on the outcome. To describe that more precisely, we need to impose some terminology and notations:

- $(X_i, Y_i), i = 1, ..., n$ are n independent random variables whereby X_i is the observation (or score) and Y_i the outcome
- c is the cutoff
- $T_i \in \{0, 1\}$ is the treatment defined by $T_i = \mathbb{1}(X_i \geq c)$

- $Y_i(1)$ and $Y_i(0)$ are the conditional outcomes in case the treatment is applied or not, respectively, whereby $Y_i = Y_i(0)(1 - T_i) + Y_i(1)T_i$
- The average treatment effect τ_Y at the cutoff is defined by

$$\tau_Y = \mathbb{E}(Y_i(1) - Y_i(0) \mid X_i = c) = \lim_{x \searrow c} \mathbb{E}(Y_i \mid X_i = x) - \lim_{x \nearrow c} \mathbb{E}(Y_i \mid X_i = x)$$

Indeed, as the average treatment effect τ_Y describes the impact of the treatment on the outcome, it is our value of interest which we want to estimate. By estimation, we try to approximate this value by the abovementioned distance of the regression functions at the cutoff. Given the above notation, we can visualize the average treatment effect as in Figure 1.

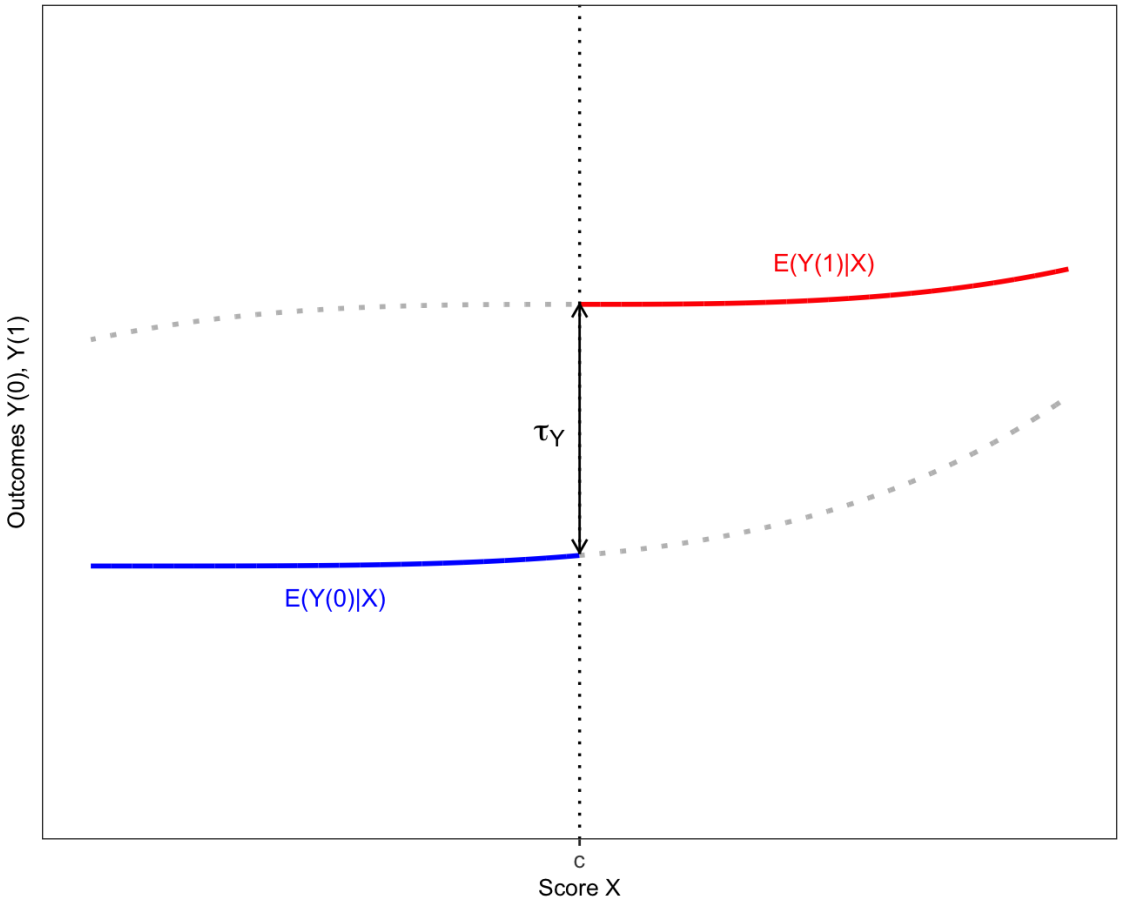


Figure 1: Average treatment effect as the jump at the cutoff (visualized by using the `R` package `ggplot2` [Wic16])

As we will see in the subsequent explanations, the estimation relies on choosing a bandwidth h, which determines the interval around the cutoff. Indeed, just the data points satisfying $X_i \in [c - h, c + h]$ are considered for estimation. The choice of that bandwidth can be quiet tricky since a too small or too big choice can lead to under- or over-smoothing. We will come back to that problem in more detail later on. However, the

estimated version of the average treatment effect via local linear regression, denoted by $\hat{\tau}_h$, is visualized in Figure 2.

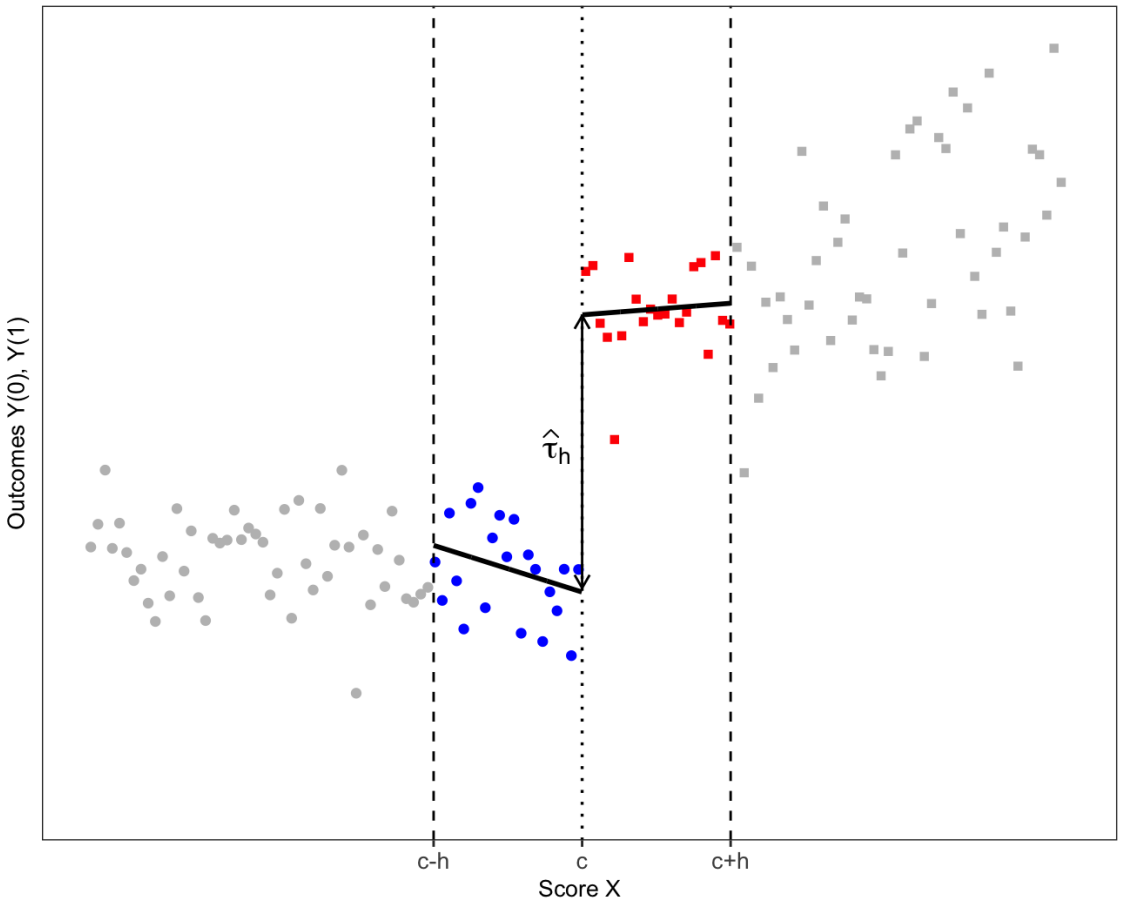


Figure 2: Estimation of the average treatment effect (visualized by using the `R` packages `rdrobust` [CCFT22] and `ggplot2` [Wic16])

As there are several small variations of the design that is described above, we will cover those within the next subsections.

3.1.1 Sharp RDD

The term *sharp* refers to the jump of the treatment probability from zero to one at the cutoff point. Indeed, Sharp RDD is exactly what we described above with the important property that $T_i = \mathbb{1}(X_i \geq c)$. Therefore, we do not need further explanation here.

3.1.2 Fuzzy RDD

In Fuzzy RDD, we have the same setup as above, including the fact that the treatment probability jumps at the cutoff. Yet, the only difference is that the jump does not necessarily has to be from zero to one. In real-world scenarios, a sharp threshold, which divides the entities into receiving or not receiving a treatment, often cannot be conducted. Rather, there still would be units receiving a treatment even though they are below the threshold

and vice versa. This situation is called *imperfect compliance*, in contrast to Sharp RDD, where we have *perfect compliance*. We can also establish a one-sided terminology: If all units with score less than the cutoff do not receive the treatment, but also some of the units with score above the cutoff do not receive the treatment, we speak of *left-sided compliance*. Vice versa, we can also define *right-sided compliance*. In both cases we can also just say *one-sided compliance*.

In order to design that behaviour, it is necessary to divide between units which are assigned to the treatment, and units which actually receive the treatment. Therefore, we need to introduce a variable T_i which indicates whether the unit is assigned to the treatment. Indeed, T_i would be defined as in the Sharp RDD case, namely $T_i = \mathbb{1}(X_i \geq c)$. In addition, we need a variable D_i which determines if a unit actually receives a treatment. There are two potential outcomes for this, $D_i(0)$, the indicator of receiving a treatment in case the unit is not assigned to it, and $D_i(1)$, the indicator of receiving a treatment in case the unit is assigned to it. Hence, $D_i = D_i(0)(1 - T_i) + D_i(1)T_i$. Figure 3 visualizes the difference of sharp RDD and fuzzy RDD in the conditional probability of receiving the treatment given the score value.

As there now can be four different possible situations (because $T_i, D_i \in \{0, 1\}$), we have four different potential outcomes for Y_i. That is $Y_i(1, D_i(1)) = D_i(1)Y_i(1, 1) + (1 - D_i(1))Y_i(1, 0)$, which either results in $Y(1, 0)$ or $Y(1, 1)$, and $Y_i(0, D_i(0)) = D_i(0)Y_i(0, 1) + (1 - D_i(0))Y_i(0, 0)$, which similarly either results in $Y(0, 0)$ or $Y(0, 1)$. Then, the observed outcome is $Y_i = T_iY_i(1, D_i(1)) + (1 - T_i)Y_i(0, D_i(0))$.

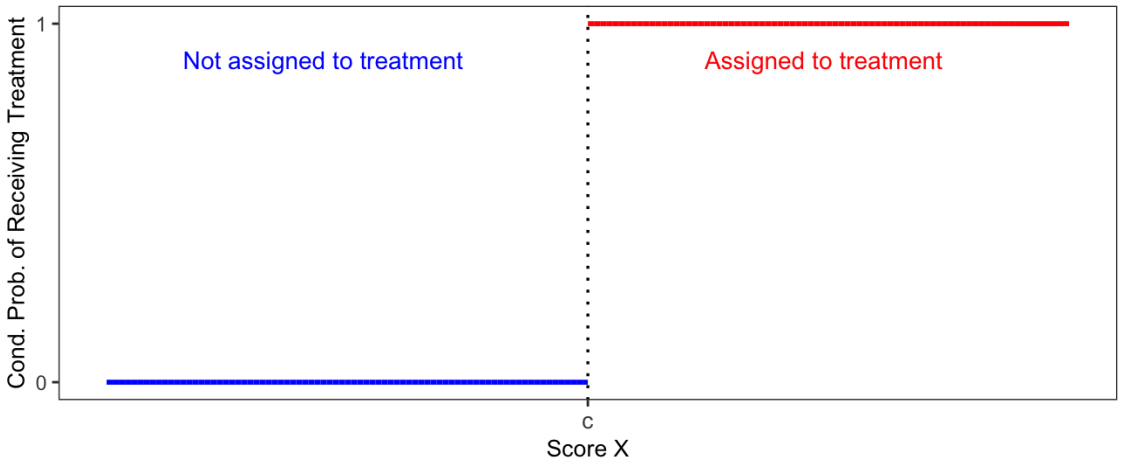
In an analogous way to Sharp RDD, we can now define the average treatment effect at the cutoff by

$$\begin{aligned}\tau_Y^{\text{fuzzy}} &= \lim_{x \searrow c} \mathbb{E}(Y_i \mid X_i = x) - \lim_{x \nearrow c} \mathbb{E}(Y_i \mid X_i = x) \\ &= \lim_{x \searrow c} \mathbb{E}(Y_i(1, D_i(1)) \mid X_i = x) - \lim_{x \nearrow c} \mathbb{E}(Y_i(0, D_i(0)) \mid X_i = x).\end{aligned}$$

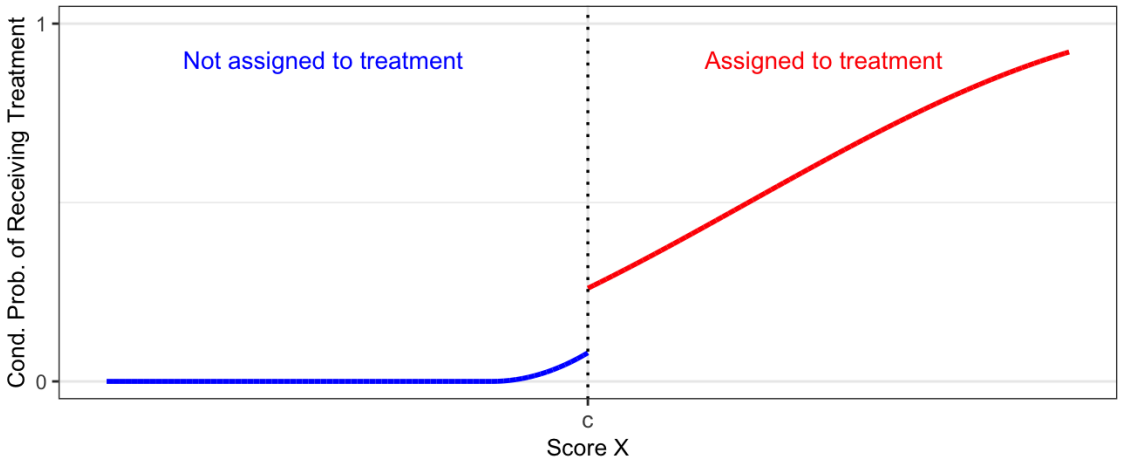
3.1.3 Kink RDD

Kink Designs cover the situation where the treatment is assigned via a continuous score formula, but the slope of the conditional expectation of the outcome given the score changes discontinuously at the cutoff point [CT21]. Then, the average treatment effect would be zero in the above manner. However, since the slope changes at the cutoff, there is still an effect of the treatment on the outcome. To describe this effect, we can examine the jump of the first derivatives of the conditional expectation of the outcome, given the score, at the cutoff point. This is visualized in Figure 4. Indeed, we describe the average treatment effect by

$$\tau_Y^{\text{kink}} = \lim_{x \searrow c} \frac{\mathrm{d}}{\mathrm{d}x} \mathbb{E}\left(Y_i \mid X_i = x\right) - \lim_{x \nearrow c} \frac{\mathrm{d}}{\mathrm{d}x} \mathbb{E}\left(Y_i \mid X_i = x\right)$$



(a) Sharp RDD (perfect compliance)



(b) Fuzzy RDD (imperfect compliance)

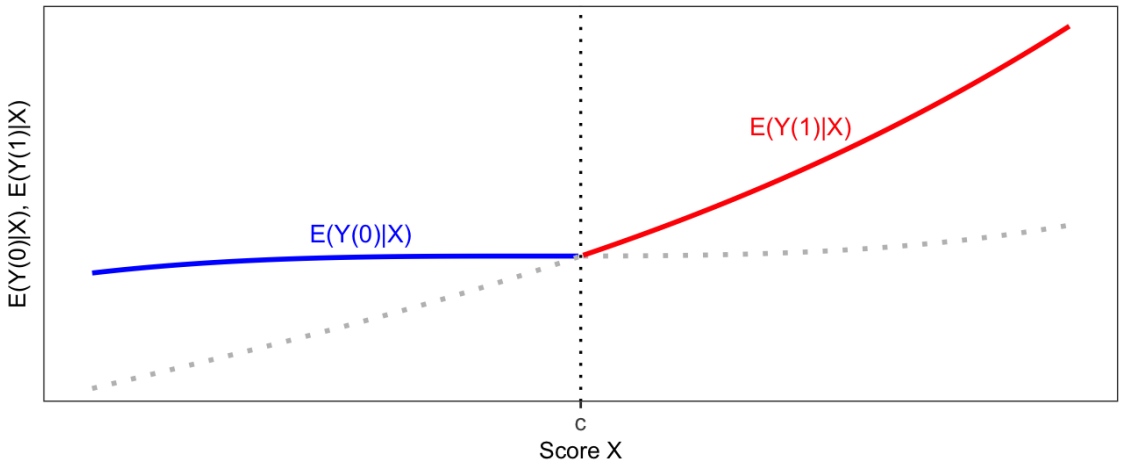
Figure 3: Comparison of the conditional probability of receiving the treatment between sharp and fuzzy RDD (visualized by using the `R` package `ggplot2` [Wic16])

which can now be estimated, e.g. by using local-quadratic estimators as suggested in [CCT14].

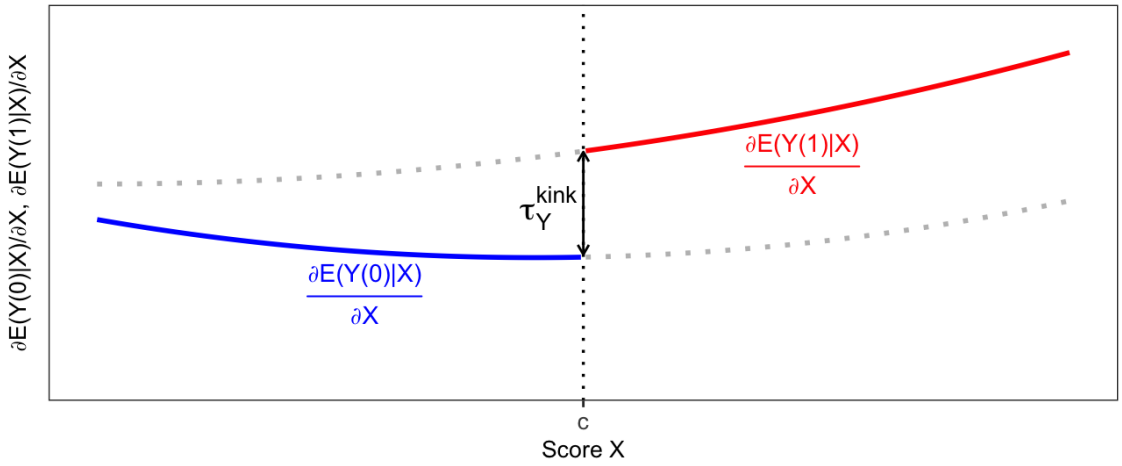
3.1.4 RDD with Multiple Cutoffs or Multiple Scores

RDD can also be modelled by using multiple cutoffs or multiple scores as described by [CTVB20]. There are three main settings:

- *Non-cumulative multiple cutoffs:* Here the units are assigned to different groups whereby each one is provided with their own univariate score. So the cutoff value is different with every group.
- *Cumulative multiple cutoffs:* All units are assigned a univariate score whereby a different treatment is applied with different score values. That means, there are multiple cutoff points which each induce its own treatment to the respective unit.



(a) Without derivative



(b) With derivative

Figure 4: Conditional expectations of potential outcomes and their derivatives at the cutoff in kink RDD (visualized by using the `R` package `ggplot2` [Wic16])

- *Multiple scores:* Units are assigned multiple scores and the treatment is assigned depending on all score values.

Modelling for non-cumulative multiple cutoffs X_i is the running variable and $(Y_i(0), Y_i(1))$ are the potential outcomes. As the new aspect is now that the cutoff can differ with each entity, the cutoff is described by a random variable $C_i \in \mathcal{C}$ with $\mathcal{C} = \{c_1, c_2, ..., c_J\}$. That means, C_i partitions the population into different groups. Then, the treatment variable T_i is described by $T_i = \mathbb{1}(X_i \geq C_i)$. In a sharp setting, we now refer to the cutoff-specific average treatment effect $\tau_Y^{\text{mc}}(c)$ for $c \in \mathcal{C}$ as

$$\tau_Y^{\text{mc}}(c) = \lim_{x \searrow c} \mathbb{E}\left(Y_i \mid X_i = x, C_i = c\right) - \lim_{x \nearrow c} \mathbb{E}\left(Y_i \mid X_i = x, C_i = c\right).$$

The pooled average treatment effect is then defined by

$$\tau_P = \lim_{x \searrow 0} \mathbb{E}\left(Y_i \mid \tilde{X}_i = x\right) - \lim_{x \nearrow 0} \mathbb{E}\left(Y_i \mid \tilde{X}_i = x\right)$$

whereby $\tilde{X}_i = X_i - C_i$ is the centered running variable. We can rewrite τ_P as

$$\tau_P = \sum_{c \in \mathcal{C}} \tau_Y^{\text{mc}}(c)\omega(c)\,, \quad \omega(c) = \frac{f_{X|C}(c \mid c)P(C_i = c)}{\sum_{c \in \mathcal{C}} f_{X|C}(c \mid c)P(C_i = c)},$$

which can now be estimated using local polynomial methods. Please refer to Figure 5 for a visualization of the average treatment effect in a setting of two non-cumulative cutoffs.

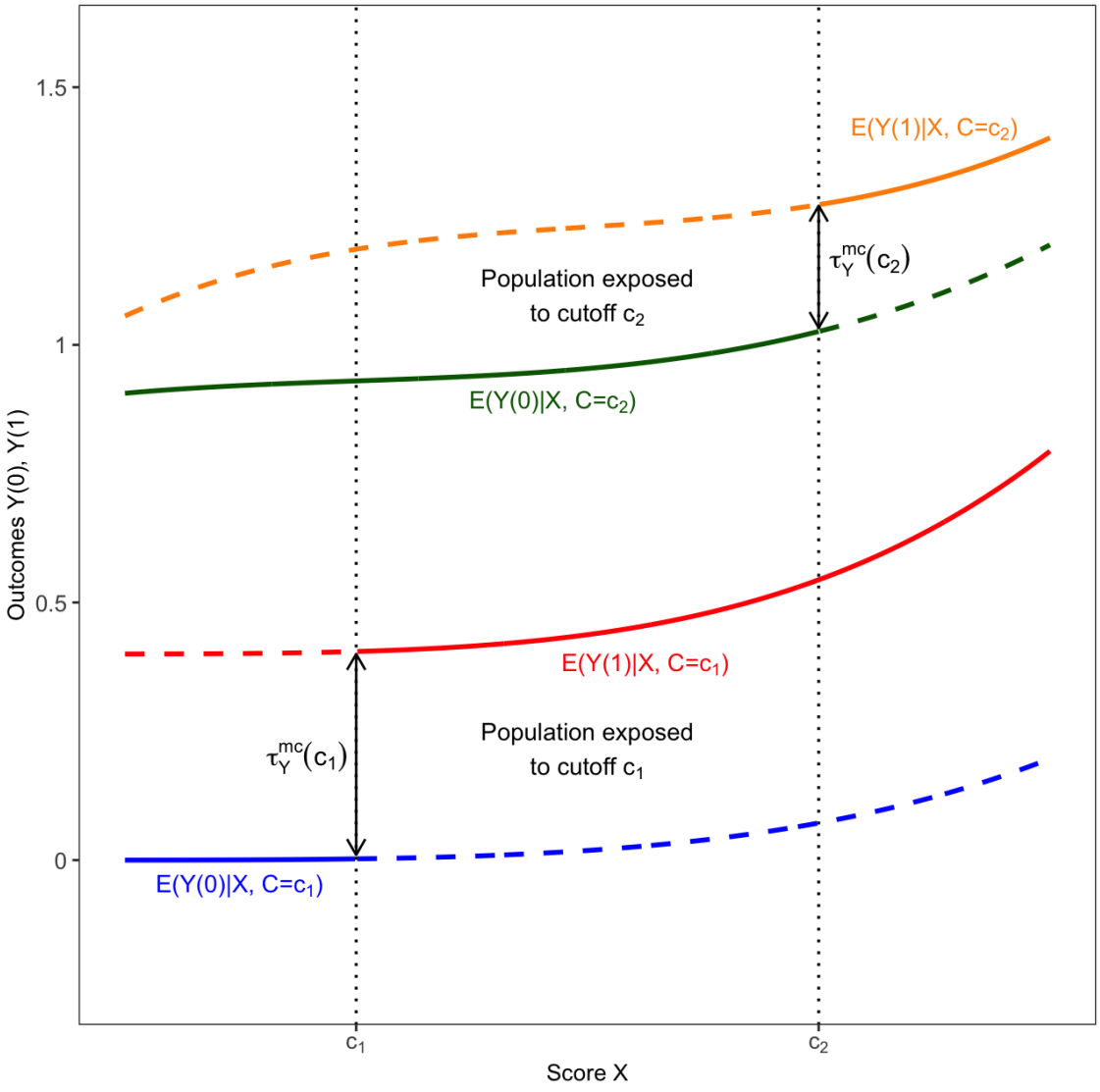


Figure 5: Average treatment effect for non-cumulative multiple cutoffs (visualized by using the `R` package `ggplot2` [Wic16])

Modelling for cumulative multiple cutoffs With different cutoffs the applied treatment (or the intensity of the treatment) changes. Therefore, we have multiple cutoff points $c_1, ..., c_J$ and treatments $T_i \in \{t_1, ..., t_{J+1}\}$ such that $T_i = t_j$ if $c_{j-1} \leq X_i \leq c_j$, whereby we set $c_0 = -\infty$ and $c_{J+1} = \infty$. Hence, we have the potential outcomes $Y(t_j), j = 1, ..., J$. Indeed, then we also have J average treatments effects $\tau_j^{\text{mc}}, j = 1, ..., J$ for each of the cutoffs, namely

$$\tau_j^{\text{mc}} = \mathbb{E}\left(Y_i(t_{j+1}) - Y_i(t_j) \mid X_i = c_j\right) = \lim_{x \searrow c_j} \mathbb{E}\left(Y_i \mid X_i = x\right) - \lim_{x \nearrow c_j} \mathbb{E}\left(Y_i \mid X_i = x\right).$$

Again, these average treatment effects can be estimated by using local polynomial methods.

Modelling for multiple scores The model given here includes two running variables since this case is studied most often in practice. It can be generalized for more than two running variables without much effort. Therefore, we consider running variables $\mathbf{X}_i = (X_{1i}, X_{2i})$. Hence, the threshold or cutoff is no longer a single value anymore, rather it is a boundary $\mathcal{B}$ of the so-called *treated area* $\mathcal{B}_t$, which we assume to be a compact set. The so-called *control region* is the area, in which the treatment is not applied, i.e. $\mathcal{B}_c = \mathbb{R}^2 \setminus \mathcal{B}_t$. As an example, we can refer to [CIT23] with according visualization in Figure 6, in which the first dimension of the running variable is the student's score in a language test and the second dimension the score in a mathematics test. We could now state that every student with a mathematics score better than 60, and a language score better than 80 receives a certain treatment. Then the boundary $\mathcal{B}$ would be

$$\mathcal{B} = \{x_1 \geq 80, x_2 = 60\} \cup \{x_1 = 80, x_2 \geq 60\}.$$

Now, we can describe the average treatment effect $\tau_Y^{\text{ms}}(\mathbf{b})$ for every point $\mathbf{b} \in \mathcal{B}$ on the boundary by

$$\tau_Y^{\text{ms}}(\mathbf{b}) = \lim_{\substack{\|\mathbf{x}-\mathbf{b}\| \to 0, \\ \mathbf{x} \in \mathcal{B}_t}} \mathbb{E}\left(Y_i \mid \mathbf{X}_i = \mathbf{x}\right) - \lim_{\substack{\|\mathbf{x}-\mathbf{b}\| \to 0, \\ \mathbf{x} \in \mathcal{B}_c}} \mathbb{E}\left(Y_i \mid \mathbf{X}_i = \mathbf{x}\right)$$

under sufficient regularity conditions in order for the limits to be well-defined.

3.2 Estimation

In general, as already mentioned above, we can distinguish two different methods for estimation. On the one hand, we have the local polynomial approach and, on the other hand, the local randomization approach. Local randomization methods require stronger assumptions than the local polynomial approach. However, this approach can be particularly useful when the local polynomial approach is not feasible due to insufficient data or for backing up the estimation by an additional result. Mainly, the randomization approach relies on two major assumptions: The first is that there needs to be a window around the cutoff in which the treatment can be assumed to be randomly assigned. Secondly, the po-

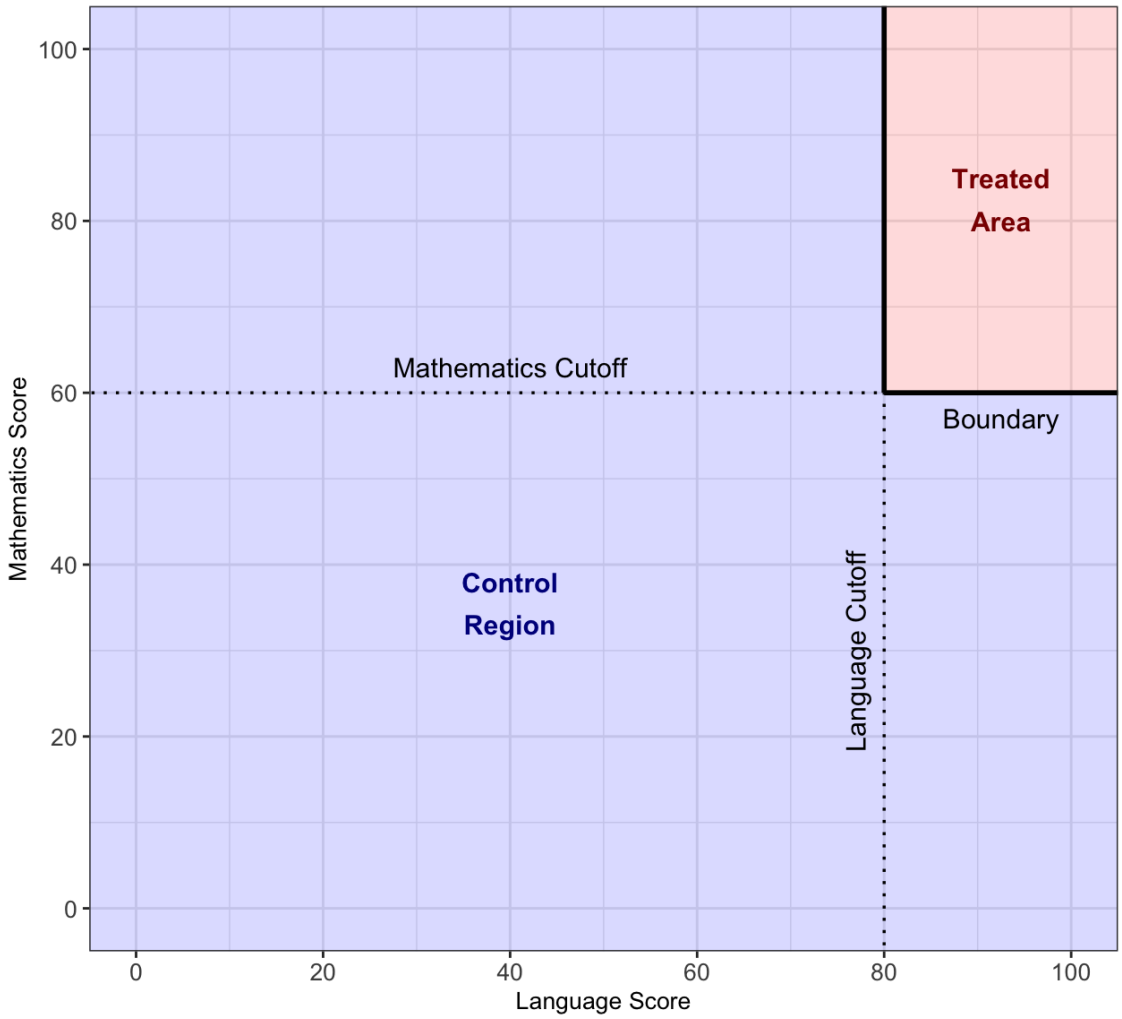


Figure 6: The average treatment effect for multiple scores (visualized by using the `R` package `ggplot2` [Wic16])

tential outcomes have to be unaffected by the score. Then, estimation can be conducted by using methods of randomized experiments. For more details and a formalization of this, we refer to [CIT23].

We want to use local linear regression for estimating the average treatment effect in a sharp RDD setting with covariates later on. Therefore, we will just focus on the average treatment effect estimation in sharp RDD settings by using the local polynomial approach as described in [CIT19]. For the study of estimation and inference in other settings and by other methods, please refer to [CIT23].

3.2.1 Overview

As already mentioned above, the average treatment effect in a sharp setting corresponds with the vertical distance between the functions $\mathbb{E}(Y_i(1) \mid X_i = x)$ and $\mathbb{E}(Y_i(0) \mid X_i = x)$ at the cutoff c. Here the problem is that data sets contain not enough information to

exactly describe these functions. Hence, we need to approximate these functions and use regression functions in order to determine the distance at the cutoff. In order that this estimator is meaningful, we need to assume continuity of $\mathbb{E}(Y_i \mid X_i = x)$ from the left and right side in c. As suggested by [CIT19], local polynomial regression, that means regression via a polynomial in a neighborhood around the cutoff, should be used to achieve a good estimator of the average treatment effect. This is done locally since a global regression is too vulnerable to extreme points that may occur far away from the cutoff, leading to a misleading estimator. In contrast to that, regression via low order polynomials in a small neighborhood around the cutoff is quiet resilient to outliers that are more far away. However, choosing the neighborhood too small can also dilute the estimator due to too few data points within the neighborhood. In the subsequent content, we will address this problem by considerations on the bandwidth choice. But first we discuss the procedure of estimation itself in more detail.

3.2.2 Procedure

As described above, local polynomial estimation takes observations near the cutoff into consideration. We do that by choosing a bandwidth $h > 0$ such that all data points within the interval $[c - h, c + h]$ are used for a polynomial regression. In fact, we want to perform the regression separately for the control and treatment area. That means, we do regressions on the respective intervals $[c - h, c]$ and $[c, c + h]$. Since points more far away from the cutoff are of less importance than points closer to the cutoff, we also want to implement a weighting scheme by setting up a kernel function. In particular, local polynomial estimation can be described with the following steps:

1. A polynomial order m of the regression and a kernel function K have to be chosen.

2. A bandwidth h has to be chosen.

3. A polynomial regression of order m for the treatment area ($c \leq X_i \leq c + h$) has to be performed: A weighted least squares regression of the observed outcome on a constant and powers of the running variable can be conducted whereby we use the kernel function K, i.e.

$$\hat{\theta}_n^+ = \underset{(\theta_0^+,\dots,\theta_m^+)\in\mathbb{R}^{m+1}}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i - c}{h}\right) \left(Y_i - \theta_0^+ - \theta_1^+ \left(\frac{X_i - c}{h}\right) - \dots - \theta_m^+ \left(\frac{X_i - c}{h}\right)^m\right)^2.$$

4. In an analogous way, we perform a local polynomial regression of order m for the

control area ($c - h \leq X_i \leq c$), i.e.

$$\hat{\theta}_n^- = \underset{(\theta_0^-,\ldots,\theta_m^-)\in\mathbb{R}^{m+1}}{\operatorname{argmin}} \sum_{i=1}^{n} K\left(\frac{X_i - c}{h}\right)\left(Y_i - \theta_0^- - \theta_1^-\left(\frac{X_i - c}{h}\right) - \ldots - \theta_m^-\left(\frac{X_i - c}{h}\right)^m\right)^2.$$

5. The first entry of $\hat{\theta}_n^+$ and $\hat{\theta}_n^-$ are estimates of the points $\mathbb{E}(Y_i(1) \mid X_i = c)$ and $\mathbb{E}(Y_i(0) \mid X_i = c)$, respectively. Therefore, we can now estimate the sharp RDD average treatment effect by

$$\hat{\tau}_h = \theta_0^+ - \theta_0^-$$

whereby $\theta_0^+ = \left(\hat{\theta}_n^+\right)^{(1)}$ and $\theta_0^- = \left(\hat{\theta}_n^-\right)^{(1)}$.

This procedure is visualized in Figure 7, whereby the order $m = 1$ for linear regression is chosen. Step 3 and 4 can also be performed in one single least squares regression via regression of Y_i on T_i, X_i and $X_i T_i$. In case of $m = 1$, that is local linear regression, it looks like

$$\hat{\theta}_n = \underset{\theta\in\mathbb{R}^4}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i - c}{h}\right)\left(Y_i - \theta^{(1)} - \theta^{(2)} T_i - \theta^{(3)}\frac{X_i - c}{h} - \theta^{(4)}\frac{T_i(X_i - c)}{h}\right)^2.$$

Then, as the second component of $\hat{\theta}_n$ describes the impact of the treatment on the outcome, we can estimate the average treatment effect by

$$\hat{\tau}_h = \hat{\theta}_n^{(2)}.$$

In the next sections, we will discuss the choice of the order m, the bandwidth h as well as the kernel K.

3.3 Kernel and Polynomial Order

Kernel functions are used to weight the importance of data points within the bandwidth interval. Typical properties of a kernel are that it is non-negative, integrable, compactly supported and integrates to one. There are three kernels that are most commonly used throughout the RDD research:

- Triangular kernel: $K : \mathbb{R} \to [0, \infty), K(x) = \max(1 - |x|, 0)$
- Uniform kernel: $K : \mathbb{R} \to [0, \infty), K(x) = \frac{1}{2} \cdot \mathbb{1}(|x| \leq 1)$
- Epanechnikov kernel: $K : \mathbb{R} \to [0, \infty), K(x) = \frac{3}{4}(1 - x^2)\mathbb{1}(|x| \leq 1)$

Since we just want to weight points within the bandwidth, we use the scaled kernel functions given by $K_h(x) = \frac{1}{h} K\left(\frac{x-c}{h}\right)$, which are visualized in Figure 8. As the choice of the kernel function does not have a strong impact on inference, each of the options would be

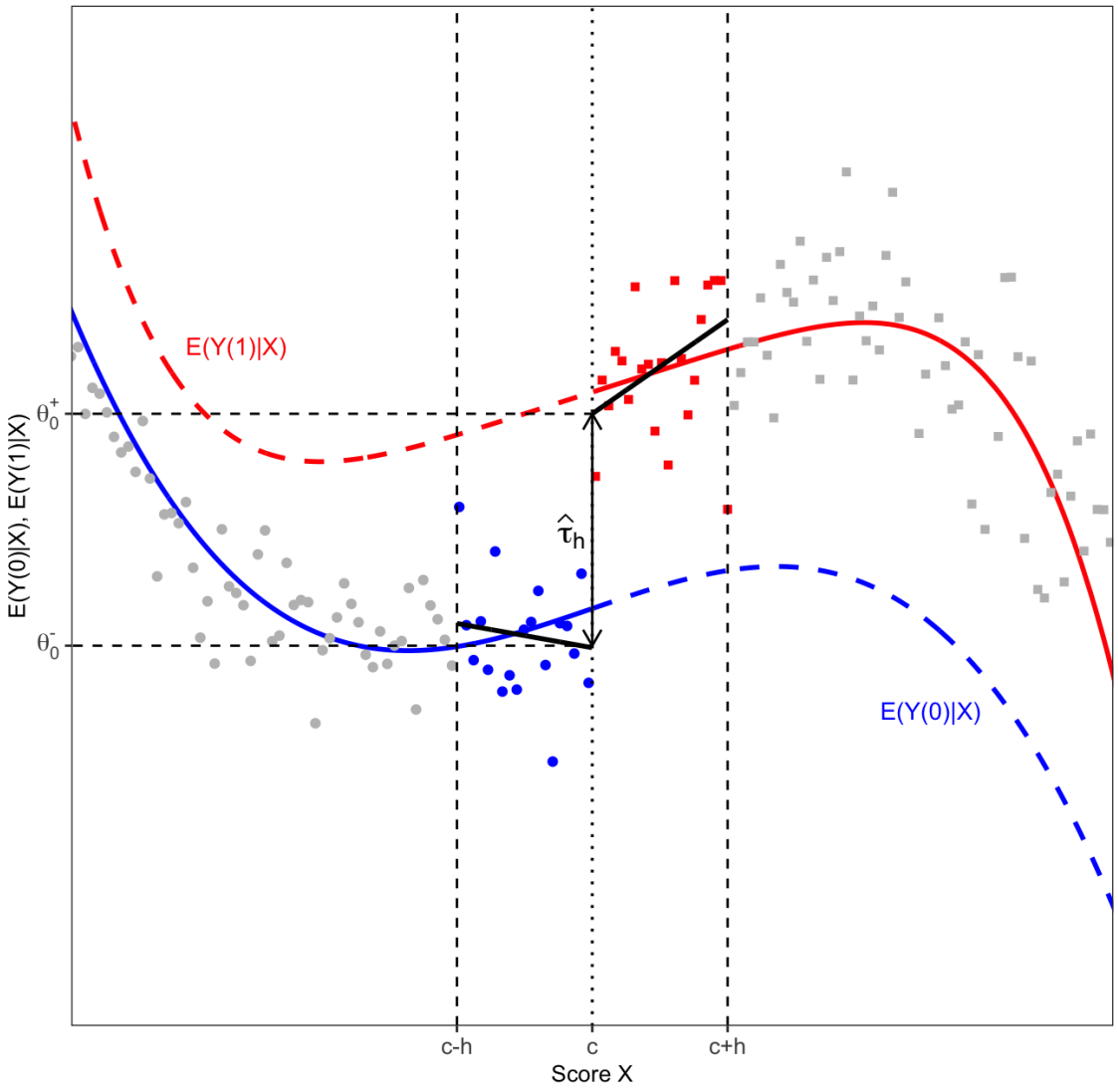


Figure 7: Local linear regression for estimation of the average treatment effect in a sharp RD design (visualized by using the `R` packages `rdrobust` [CCFT22] and `ggplot2` [Wic16])

an appropriate choice. Nonetheless, [CIT19] suggests using the triangular kernel as this implies an estimator with optimal properties when used with a MSE-optimized bandwidth.

As written above, we also need to choose the degree of the regression polynomial. First of all, a degree of zero, that is a constant regression function, is not very beneficial since it can cause huge errors at the boundary points which are our points of interest in RDD. Therefore, we need to choose a degree of at least one. Of course, given a fixed bandwidth, choosing a larger degree makes the regression more accurate. However, a larger degree also implies stronger variability of the treatment effect estimator. Especially the values at the boundary points may tend to be more unreliable. Hence, linear regression is most commonly used in RDD research [CIT19].

Even more important is the appropriate choice of the bandwidth since this determines

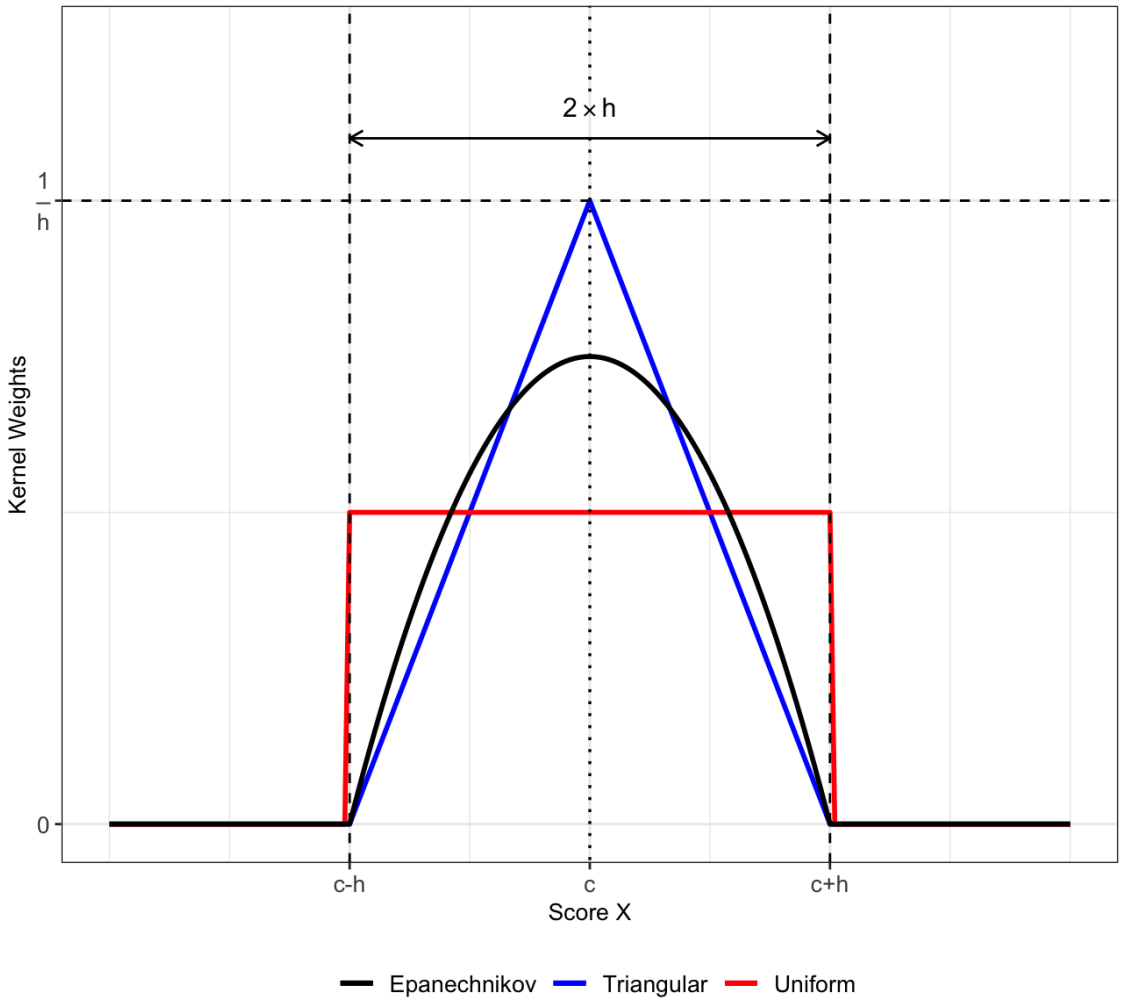


Figure 8: Different kernel functions as weighting scheme for the local regression (visualized by using the `R` package `ggplot2` [Wic16])

whether our local linear regression delivers useful and reliable results or not. This will be discussed next.

3.4 Choice of Bandwidth

We start by motivating the discussion on the bandwidth choice. Figure 9 illustrates the estimation of the average treatment effect via local linear regression for two different choices of the bandwidth. It suggests that different bandwidth choices can lead to different regression errors at the cutoff and therefore to a varying precision of our estimator. We can see in the graph that bandwidth h_1 leads to a much better approximation than h_2 since the function is almost linear in this interval around the cutoff. Indeed, decreasing the bandwidth for a fixed polynomial order will lower the bias, that is the error of the estimator. However, we have to pay attention since lowering the bandwidth also causes less observations that fall inside the bandwidth area leading to an increasing variance of

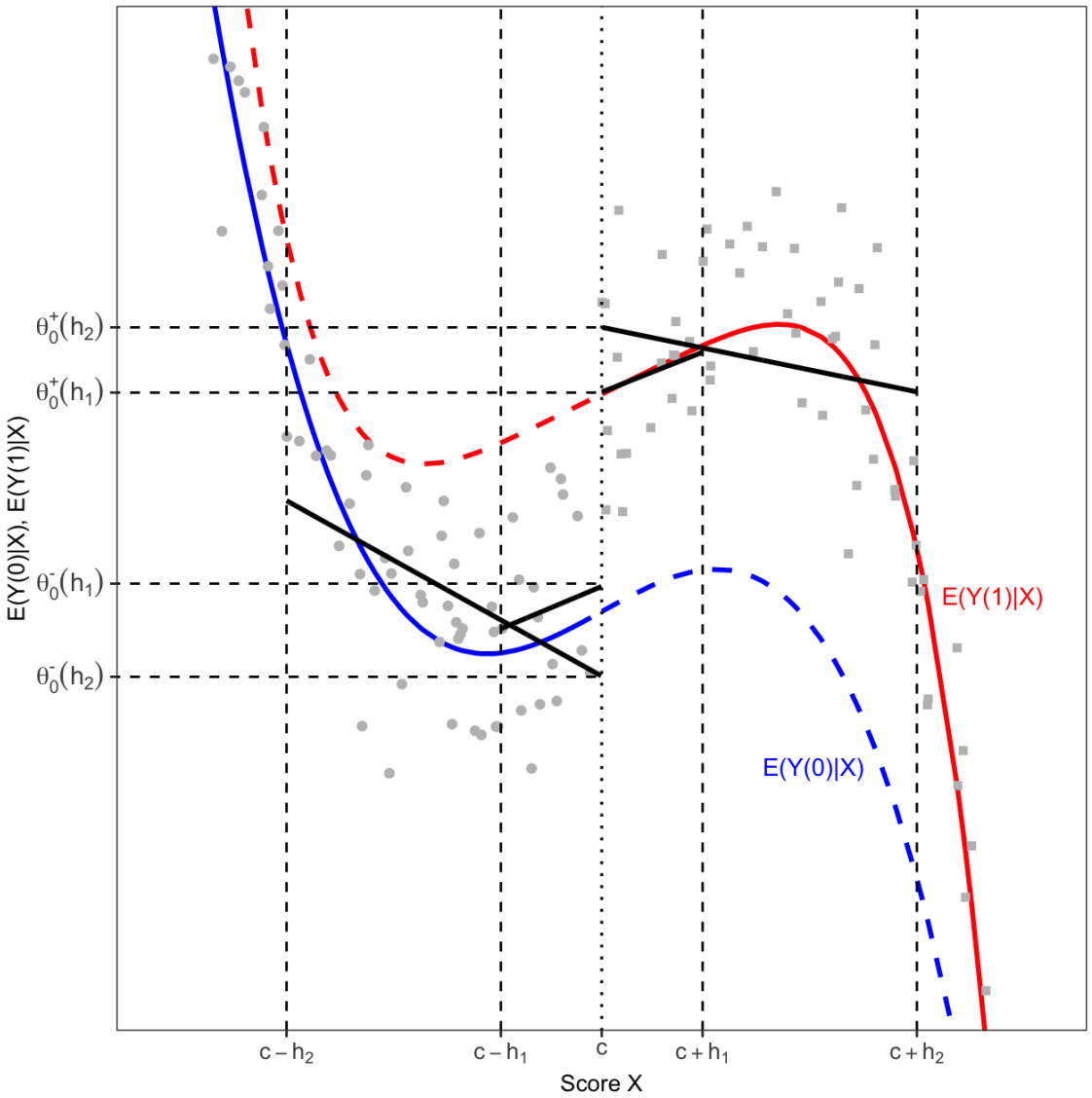


Figure 9: Bandwidth comparison for estimation by local linear regression (visualized by using the `R` packages `rdrobust` [CCFT22] and `ggplot2` [Wic16])

our estimator. This is why we cannot just lower the bandwidth as much as we want in order to decrease the bias. Rather, we have to find a balance between bias and variance, which is also called *bias-variance trade-off*. Therefore, the most popular approach is to minimize the mean squared error

$$\mathrm{MSE}(\hat{\tau}_h) = \mathrm{Bias}^2(\hat{\tau}_h) + \mathrm{Var}(\hat{\tau}_h).$$

MSE-optimal bandwidth In order to determine a bandwidth which optimizes the MSE, we have to find expressions for both bias and variance of the estimator. We want to omit the details here since the representations will be shown in the more general case when covariates are included. Yet, we want to present the main statements in order to

be able to compare the results of the setting with and without covariates, and to examine if covariates actually enhance the precision of our estimator. The approximate bias and variance can be described by

$$\text{Bias}(\hat{\tau}_h) = h^{2(m+1)}\mathcal{B} \quad \text{and} \quad \text{Var}(\hat{\tau}_h) = \frac{1}{nh}\mathcal{V}$$

whereby $\mathcal{B}$ and $\mathcal{V}$ are constants which do not depend on the sample size and the bandwidth choice. We have

$$\mathcal{B} = \mathcal{B}_+ - \mathcal{B}_-, \quad \mathcal{B}_- \approx \mu_-^{(m+1)} B_-, \quad \mathcal{B}_+ \approx \mu_+^{(m+1)} B_+ \tag{3.1}$$

with $\mu_-^{(k)} = \lim_{x \nearrow c} \frac{\mathrm{d}^k}{\mathrm{d}x^k}\mathbb{E}(Y_i(0) \mid X_i = x)$ and $\mu_+^{(k)} = \lim_{x \searrow c} \frac{\mathrm{d}^k}{\mathrm{d}x^k}\mathbb{E}(Y_i(1) \mid X_i = x)$ for some constants B_- and B_+ depending on the kernel function and the order of the polynomial used for regression. Moreover,

$$\mathcal{V} = \mathcal{V}_+ + \mathcal{V}_-, \quad \mathcal{V}_- \approx \frac{\sigma_-^2}{f} V_-, \quad \mathcal{V}_- \approx \frac{\sigma_+^2}{f} V_+ \tag{3.2}$$

whereby

$$\sigma_-^2 = \lim_{x \nearrow c} \text{Var}(Y_i(0) \mid X_i = x) \quad \text{and} \quad \sigma_+^2 = \lim_{x \searrow c} \text{Var}(Y_i(1) \mid X_i = x)$$

for some constants V_- and V_+ depending on the kernel and the polynomial regression order, and f describing the density of the score at the cutoff. Finding the MSE-optimal bandwidth is now reduced to solving the minimization problem

$$\min_{h>0} \left(h^{2(m+1)}\mathcal{B}^2 + \frac{1}{nh}\mathcal{V} \right).$$

Doing that, we obtain

$$h_{\text{MSE}} = \left(\frac{\mathcal{V}}{2(m+1)\mathcal{B}^2} \right)^{\frac{1}{2m+3}} n^{-\frac{1}{2m+3}}$$

which is the MSE-optimal bandwidth. In practice, as $\mathcal{V}$ and $\mathcal{B}$ are unknown, this equation is infeasible to solve. Therefore, these variables need to be estimated, which is for example described in [CIT19] and [IK12]. We do not go into more detail here and rather discuss next, how our estimator can be improved by including additional covariates.

3.5 Including Covariates

Additionally to the above described RDD, we could come up with the question if additional information about the units increases the precision of our estimator and helps to determine the impact of the treatment on the outcome. So let us assume that we have covariates $Z_i \in \mathbb{R}^p, i = 1, ..., n$, whose information we also want to incorporate into our regression.

As described in [CCFT19], there are several ways to do that, which all result in slightly different assumptions on the covariates. Here, we just want to focus on one method which includes the covariates in the local linear regression linearly. Indeed, we naturally extend the regression without covariates, that is

$$\hat{Y}_i = \hat{\theta}_1 + \hat{\theta}_2 T_i + \hat{\theta}_3 \frac{X_i - c}{h} + \hat{\theta}_4 \frac{T_i(X_i - c)}{h},$$

to a regression of the form

$$\tilde{Y}_i = \tilde{\theta}_1 + \tilde{\theta}_2 T_i + \tilde{\theta}_3 \frac{X_i - c}{h} + \tilde{\theta}_4 \frac{T_i(X_i - c)}{h} + Z_i^\top \tilde{\gamma}$$

with $\tilde{\gamma} \in \mathbb{R}^p$. Then, $\tilde{\theta}_2$ is an estimator for the average treatment effect including covariates, which we also call *covariate-adjusted estimator*. In order that this estimator is consistent, we need to impose assumptions that contain, on the one hand, the well-known continuity assumptions of RDD without covariates and, on the other hand, additional assumptions regarding the covariates. One of the main requirements is that the covariates need to be *pre-treatment covariates*, which means that the treatment has no impact on the covariates. We can formalize this by the key assumption

$$\mu_{Z+} = \mu_{Z-},$$

whereby $\mu_{Z-} = \lim_{x \nearrow c} \mathbb{E}(Z_i \mid X_i = x)$ and $\mu_{Z+} = \lim_{x \searrow c} \mathbb{E}(Z_i \mid X_i = x)$, which particularly states that $\mathbb{E}(Z_i \mid X_i = x)$ has no jump at the cutoff.

Mostly, the covariate-adjusted estimator is more efficient than the baseline estimator without considering covariates. Nonetheless, particularly if the number of covariates is relatively high in comparison to the number of observations, the estimator lacks efficiency and provides a misleading approximation in settings with finitely many samples. This is due to overfitting, which is caused by too many unknown parameters diluting the result. Especially, when the number of covariates exceeds the number of observations, the above estimator $\tilde{\theta}_2$ is not even uniquely defined. For this reason, the covariate-adjusted estimator of the average treatment effect is just considered to be appropriate in low-dimensional settings [KR21].

As mentioned in the introduction, the first main goal of this thesis is to prove that this estimator is consistent and asymptotically normally distributed under sufficient regularity conditions. This means we want to show that

$$\frac{\sqrt{nh}\left(\hat{\tau}_h - \tau_Y - h^2 \mathcal{B}_n\right)}{\mathcal{S}_n} \xrightarrow{d} \mathcal{N}(0,1)$$

whereby $\hat{\tau}_h$ is the covariate-adjusted estimator, τ_Y the average treatment effect, $\mathcal{B}_n$ the bias and $\mathcal{S}_n$ the variance. This will be proved in Chapter 7, in which we will also give explicit representations of the bias and the variance and discuss the improvement of those caused by including covariates. Of course, as multiple assumptions are necessary in order

that the above statement holds, they will also be given within this chapter.

Chapter 4 until 6 will provide us will all the necessary foundations in order to be able to prove the main theorem. In particular, the next chapter will formally impose all the required terminology and notation as we define our RDD setup with covariates used in the remainder of this thesis.

4 Setup and Notation

The data is given by an independent sample $\{(Y_i, X_i, Z_i), i = 1, ..., n\}$ whereby n is the population size, $Y_i \in \mathbb{R}$ is the outcome variable, $X_i \in \mathbb{R}$ the running variable with its probability density function f_X and $Z_i \in \mathbb{R}^p$ is a vector of pre-treatment covariates. The running variable X_i decides whether a unit receives a treatment or not. In fact, we define the treatment variable T_i by $T_i := \mathbb{1}(X_i \geq c)$, which means that a treatment is applied if and only if the running variable exceeds a certain cutoff c. Without loss of generality and for the sake of simplicity we want this cutoff to be zero throughout the remaining content, that is $c = 0$ and $T_i = \mathbb{1}(X_i \geq 0)$. Also, we have the potential outcomes $Y_i(0)$ and $Y_i(1)$, which are random variables for the outcome in case a unit receives a treatment or not, respectively. Therefore, we have that $Y_i = Y_i(T_i)$, which is the real outcome depending on the value of the treatment variable.

Since we are interested in the impact of the treatment on the outcome variable, we want to examine the average treatment effect

$$\tau_Y = \mathbb{E}(Y_i(1) - Y_i(0) \mid X_i = 0).$$

As we will always be in the situation that $\mathbb{E}(Y_i \mid X_i = x)$ is left- and right-continuous in zero, we can also determine it by

$$\tau_Y = \lim_{x \searrow 0} \mathbb{E}(Y_i \mid X_i = x) - \lim_{x \nearrow 0} \mathbb{E}(Y_i \mid X_i = x).$$

For arbitrary random variables A and B, we want to define $\mu_A(x) = \mathbb{E}(A \mid X_i = x)$, $\sigma^2_{AB}(x) = \mu_{AB^\top}(x) - \mu_A(x)\mu_B(x)^\top$ and $\sigma^2_A(x) = \sigma^2_{AA}(x)$. Moreover, given a function f for which the left- and right-sided limit in zero exists, we set $f_+ = \lim_{x \searrow 0} f(x)$ and $f_- = \lim_{x \nearrow 0} f(x)$. Then we can also represent the average treatment effect by

$$\tau_Y = \mu_{Y+} - \mu_{Y-}.$$

Indeed, this is the value that we want to estimate by a consistent estimator. We will do that by using the local linear adjustment estimator

$$\hat{\tau}_h = e_2^\top \left(\hat{\theta}_n, \hat{\gamma}_n\right),$$

whereby

$$\left(\hat{\theta}_n, \hat{\gamma}_n\right) = \underset{(\theta,\gamma) \in \mathbb{R}^{p+4}}{\operatorname{argmin}} \sum_{i=1}^{n} K_h(X_i) \left(Y_i - V_i^\top \theta - Z_i^\top \gamma\right)^2.$$

Here K is a kernel function,

$$K_h(x) := \frac{1}{h} K\left(\frac{x}{h}\right),$$

$h > 0$ the bandwidth and

$$V_i := \begin{pmatrix} 1 \\ T_i \\ X_i/h \\ T_i X_i/h \end{pmatrix}.$$

For the sake of notational simplicity we will just write V_i instead of $V_i(h)$ and omit its dependency of h. However, it should be noted that this dependency exists and also needs to be considered when looking at the convergence of expressions containing V_i. Moreover, if we use the term *kernel* in the remainder of this thesis, we refer to a function which is defined as follows.

Definition 4.1 (Kernel function)**.** *A kernel is a function $f : \mathbb{R} \to \mathbb{R}$ which*

- *is non-negative and integrable,*
- *integrates to one, i.e. $\int_{-\infty}^{\infty} K(x)\,dx = 1$ and*
- *is symmetric, i.e. $K(-x) = K(x)$.*

In order to proof the asymptotic normality later on, we will have to deal with the regression error

$$r_i(h) = Y_i - V_i^\top \theta_0(h) - Z_i^\top \gamma_0(h)$$

whereby $\theta_0(h)$ and $\gamma_0(h)$ are the population regression coefficients defined by

$$\big(\theta_0(h), \gamma_0(h)\big) = \underset{(\theta,\gamma)\in\mathbb{R}^{p+4}}{\operatorname{argmin}} \; \mathbb{E}\left(K_h(X_i) \left(Y_i - V_i^\top \theta - Z_i^\top \gamma \right)^2 \right).$$

The goal of the next chapter is to show that

$$\theta_0^{(2)} = \tau_Y + h^2 \mathcal{B}_n$$

whereby we will derive an explicit representation for the bias $\mathcal{B}_n$.

5 Computing the Bias

The aim of this section is to provide a representation of the bias. This will be achieved by using a kernel matrix and performing numerous Taylor expansions. We do this by adapting the approach of [KR21] to the case of a fixed number of covariates. Overall, this chapter builds the foundation for the first step of the proof of asymptotic normality of our average treatment effect estimator.

We start with a simple definition, before we begin with the proof of invertibility of the kernel matrix.

Definition 5.1. *Let $L : \mathbb{R} \to \mathbb{R}$ be an arbitrary integrable function. Then we define*

$$L_-^{(\alpha)} = \int_{-\infty}^{0} L(u)u^\alpha \, du, \quad L_+^{(\alpha)} = \int_0^{\infty} L(u)u^\alpha \, du \quad \textit{and} \quad L^{(\alpha)} = \int_{-\infty}^{\infty} L(u)u^\alpha \, du.$$

Please recall that a kernel function, in particular, is symmetric and integrates to one by our definition.

Lemma 5.1. *Let K be a kernel with $K^{(2)} < \infty$. Then the matrix*

$$\kappa(K) := \begin{pmatrix} K^{(0)} & K_+^{(0)} & K^{(1)} & K_+^{(1)} \\ K_+^{(0)} & K_+^{(0)} & K_+^{(1)} & K_+^{(1)} \\ K^{(1)} & K_+^{(1)} & K^{(2)} & K_+^{(2)} \\ K_+^{(1)} & K_+^{(1)} & K_+^{(2)} & K_+^{(2)} \end{pmatrix}$$

is invertible.

Proof. We prove that the determinant of the matrix is unequal to zero. The fact that

$$K^{(0)} = 1, K_+^{(0)} = K_-^{(0)} = \frac{1}{2}, K_+^{(1)} = -K_-^{(1)}, K^{(1)} = 0 \text{ and } K_+^{(2)} = K_-^{(2)} \tag{5.1}$$

will be useful in the following calculation:

$$\begin{aligned} &\det \kappa(K) \\ &= 1 \cdot \det \begin{pmatrix} \frac{1}{2} & K_+^{(1)} & K_+^{(1)} \\ K_+^{(1)} & 2K_+^{(2)} & K_+^{(2)} \\ K_+^{(1)} & K_+^{(2)} & K_+^{(2)} \end{pmatrix} - \frac{1}{2} \det \begin{pmatrix} \frac{1}{2} & 0 & K_+^{(1)} \\ K_+^{(1)} & 2K_+^{(2)} & K_+^{(2)} \\ K_+^{(1)} & K_+^{(2)} & K_+^{(2)} \end{pmatrix} \\ &\quad + 0 \det \begin{pmatrix} \frac{1}{2} & 0 & K_+^{(1)} \\ \frac{1}{2} & K_+^{(1)} & K_+^{(1)} \\ K_+^{(1)} & K_+^{(2)} & K_+^{(2)} \end{pmatrix} - K_+^{(1)} \det \begin{pmatrix} \frac{1}{2} & 0 & K_+^{(1)} \\ \frac{1}{2} & K_+^{(1)} & K_+^{(1)} \\ K_+^{(1)} & 2K_+^{(2)} & K_+^{(2)} \end{pmatrix} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}(2K_+^{(2)}K_+^{(2)} - (K_+^{(2)})^2) - K_+^{(1)}(\underbrace{K_+^{(1)}K_+^{(2)} - K_+^{(1)}K_+^{(2)}}_{=0}) + K_+^{(1)}(K_+^{(1)}K_+^{(2)} - 2K_+^{(1)}K_+^{(2)}) \\
&\quad - \frac{1}{2}\cdot\frac{1}{2}(2K_+^{(2)}K_+^{(2)} - (K_+^{(2)})^2) - \frac{1}{2}K_+^{(1)}K_+^{(2)}K_+^{(1)} + K_+^{(1)}K_+^{(1)}K_+^{(2)} \\
&\quad - K_+^{(1)}\cdot\frac{1}{2}(K_+^{(1)}K_+^{(2)} - 2K_+^{(1)}K_+^{(2)}) - K_+^{(1)}K_+^{(1)}K_+^{(2)} + (K_+^{(1)})^2(K_+^{(1)})^2 \\
&= \frac{1}{4}(K_+^{(2)})^2 - (K_+^{(1)})^2K_+^{(2)} + (K_+^{(1)})^4 \\
&= \left((K_+^{(1)})^2 - \frac{1}{2}K_+^{(2)}\right)^2.
\end{aligned}
$$

This expression is not zero since

$$
\begin{aligned}
\left(K_+^{(1)}\right)^2 = \left(\int_0^\infty K(u)u\,du\right)^2 &= \left|\int_0^\infty K(u)u\,du\right|^2 \\
= \left(\int_0^\infty K(u)|u|\,du\right)^2 &= \left(\frac{1}{2}\int_{-\infty}^\infty K(u)|u|\,du\right)^2 \\
\overset{(*)}{<} \frac{1}{4}\int_{-\infty}^\infty K(u)u^2\,du &= \frac{1}{4}K^{(2)} = \frac{1}{2}K_+^{(2)},
\end{aligned}
$$

whereby inequality $(*)$ is true due to Jensen's inequality, which is stated in Proposition 1. In particular, since it is applied on a non-linear function (namely quadratic function), the implied inequality is strict. Note that we could apply Jensen's inequality as our imposed assumptions make K a probability density function. Overall, this proves the invertibility of $\kappa(K)$. □

Next, we will define a term which is highly relevant for the remainder of this thesis.

Definition 5.2. *We call a function $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ k-times one-sided differentiable at 0 when*

- *f can be extended continuously to zero from the left and from the right,*
- *all derivatives of f up to order k exist on $(-\infty, 0)$ and $(0, \infty)$ and can be extended continuously to zero from the left and from the right.*

Then we denote the value of the continuous extension of f to zero from the left by f_- and from the right by f_+. In an analogous manner, we use that notation for the extension of the derivatives, such as f'_- or f'_+.

Please recall that the probability density function of X_i is denoted by f_X. For the subsequent considerations we need the following basic statement: Let L be an integrable function and $f : \mathbb{R} \to \mathbb{R}$ be a twice one-sided differentiable function at zero. Also let f_X be twice continuously differentiable. Then, we can perform a Taylor expansion of f at zero respectively from the left and from the right by evaluating the function at zero by using the continuous extension. This leads to the Taylor expansion of $L(u)f(uh)f_X(uh)$

at $h = 0$ from the left, namely

$$\begin{aligned}
&L(u)f(uh)f_X(uh) \\
&= L(u)f_- f_X(0) + L(u)\lim_{x\nearrow 0}\frac{\mathrm{d}}{\mathrm{d}x}(f(ux)f_X(ux))h + \frac{1}{2}L(u)\lim_{x\nearrow 0}\frac{\mathrm{d}^2}{\mathrm{d}x^2}(f(ux)f_X(ux))h^2 + o(h^2) \\
&= L(u)f_- f_X(0) + hL(u)u(f\cdot f_X)'_- + \frac{h^2}{2}L(u)u^2(f\cdot f_X)''_- + o(h^2)
\end{aligned}$$

for $h \to 0$, and in an analogous way from the right:

$$\begin{aligned}
&L(u)f(uh)f_X(uh) \\
&= L(u)f_+ f_X(0) + hL(u)u(f\cdot f_X)'_+ + \frac{h^2}{2}L(u)u^2(f\cdot f_X)''_+ + o(h^2)
\end{aligned}$$

for $h \to 0$. As a consequence, we obtain under the additional assumption $L^{(2)} < \infty$ that

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{h}L\left(\frac{X_i}{h}\right)f(X_i)\right) \\
&= \int_{-\infty}^{0}\frac{1}{h}L\left(\frac{x}{h}\right)f(x)f_X(x)\,dx + \int_0^{\infty}\frac{1}{h}L\left(\frac{x}{h}\right)f(x)f_X(x)\,dx \\
&= \int_{-\infty}^{0}L(u)f(uh)f_X(uh)\,du + \int_0^{\infty}L(u)f(uh)f_X(uh)\,du \\
&= \int_{-\infty}^{0}\left[L(u)f_-f_X(0) + hL(u)u(f\cdot f_X)'_- + \frac{h^2}{2}L(u)u^2(f\cdot f_X)''_- + o(h^2)\right]du \\
&\quad + \int_0^{\infty}\left[L(u)f_+f_X(0) + hL(u)u(f\cdot f_X)'_+ + \frac{h^2}{2}L(u)u^2(f\cdot f_X)''_+ + o(h^2)\right]du \\
&= L^{(0)}_- f_- f_X(0) + L^{(0)}_+ f_+ f_X(0) + h\left[L^{(1)}_-(f\cdot f_X)'_- + L^{(1)}_+(f\cdot f_X)'_+\right] \\
&\quad + \frac{h^2}{2}\left[L^{(2)}_-(f\cdot f_X)''_- + L^{(2)}_+(f\cdot f_X)''_+\right] + o(h^2)
\end{aligned} \tag{5.2}$$

for $h \to 0$. Please note that we used $\int o(h^2)\,du = o(h^2)$ which can be verified by taking a representation of the remainder term of the Taylor expansion (e.g. the Peano form) and using the Theorem of Dominated Convergence to move the limit inside the integral.

In an analogous way, if we adjust our assumptions accordingly up to order three, we can also state

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{h}L\left(\frac{X_i}{h}\right)f(X_i)\right) \\
&= L^{(0)}_- f_- f_X(0) + L^{(0)}_+ f_+ f_X(0) + h\left[L^{(1)}_-(f\cdot f_X)'_- + L^{(1)}_+(f\cdot f_X)'_+\right] \\
&\quad + \frac{h^2}{2}\left[L^{(2)}_-(f\cdot f_X)''_- + L^{(2)}_+(f\cdot f_X)''_+\right] + O(h^3).
\end{aligned} \tag{5.3}$$

Definition 5.3 (Average treatment effect)**.** *Let $m \in \mathbb{N}$ and $A \in \mathbb{R}^m$ be a random variable such that the left- and right-sided limit of μ_A in zero exist. Then we define the average treatment effect of T_i on A as*

$$\tau_A = \mu_{A+} - \mu_{A-}.$$

Lemma 5.2. *Let f_X be three times continuously differentiable in a neighborhood around zero and A a random variable such that the function $\mu_A = \mathbb{E}(A \mid X_i = x)$ is well-defined and three times one-sided differentiable at 0. Moreover, let K be a kernel with $K^{(4)} < \infty$. As we set*

$$B(K, A) := \frac{1}{2}\kappa(K)^{-1} \left[\begin{pmatrix} K_+^{(2)} \\ K_+^{(2)} \\ K_+^{(3)} \\ K_+^{(3)} \end{pmatrix} [\mu_A f_X]''_+ + \begin{pmatrix} K_-^{(2)} \\ 0 \\ K_-^{(3)} \\ 0 \end{pmatrix} [\mu_A f_X]''_- \right],$$

it holds that

$$\kappa(K)^{-1}\mathbb{E}(K_h(X_i)V_iA) = \begin{pmatrix} f_X(0)\mu_{A-} \\ f_X(0)\tau_A \\ h[\mu_A f_X]'_- \\ h([\mu_A f_X]'_+ - [\mu_A f_X]'_-) \end{pmatrix} + h^2 B(K, A) + O(h^3)$$

for $h \to 0$.

Proof. We have

$$\begin{aligned} \mathbb{E}(K_h(X_i)V_iA) &= \mathbb{E}(\mathbb{E}(K_h(X_i)V_iA \mid X_i)) \\ &= \mathbb{E}(K_h(X_i)V_i\mathbb{E}(A \mid X_i)) \\ &= \begin{pmatrix} \mathbb{E}(K_h(X_i)\mathbb{E}(A \mid X_i)) \\ \mathbb{E}(K_h(X_i)T_i\mathbb{E}(A \mid X_i)) \\ \mathbb{E}(K_h(X_i)\frac{X_i}{h}\mathbb{E}(A \mid X_i)) \\ \mathbb{E}(K_h(X_i)\frac{T_iX_i}{h}\mathbb{E}(A \mid X_i)) \end{pmatrix} \end{aligned} \tag{5.4}$$

We perform a Taylor expansion for each of the components now. In fact, we have:

- Set $L(u) = K(u)$ and $f(x) = \mu_A(x)$ in (5.3) to obtain

$$\begin{aligned} \mathbb{E}(K_h(X_i)\mathbb{E}(A \mid X_i)) = K_-^{(0)}\mu_{A-}f_X(0) + K_+^{(0)}\mu_{A+}f_X(0) \\ + h\left(K_-^{(1)}[\mu_A f_X]'_- + K_+^{(1)}[\mu_A f_X]'_+\right) \\ + \frac{h^2}{2}\left(K_-^{(2)}[\mu_A f_X]''_- + K_+^{(2)}[\mu_A f_X]''_+\right) \\ + O(h^3). \end{aligned}$$

- Set $L(u) = K(u)\mathbb{1}(u \geq 0)$ and $f(x) = \mu_A(x)$ in (5.3) to obtain

$$\mathbb{E}(K_h(X_i)T_i\mathbb{E}(A \mid X_i)) = K_+^{(0)}\mu_{A+}f_X(0) + hK_+^{(1)}[\mu_A f_X]'_+ + \frac{h^2}{2}K_+^{(2)}[\mu_A f_X]''_+ + O(h^3).$$

- Set $L(u) = K(u)u$ and $f(x) = \mu_A(x)$ in (5.3) to obtain

$$\begin{aligned}\mathbb{E}\left(K_h(X_i)\frac{X_i}{h}\mathbb{E}(A \mid X_i)\right) = {} & K_-^{(1)}\mu_{A-}f_X(0) + K_+^{(1)}\mu_{A+}f_X(0) \\ & + h\left(K_-^{(2)}[\mu_A f_X]'_- + K_+^{(2)}[\mu_A f_X]'_+\right) \\ & + \frac{h^2}{2}\left(K_-^{(3)}[\mu_A f_X]''_- + K_+^{(3)}[\mu_A f_X]''_+\right) \\ & + O(h^3).\end{aligned}$$

- Set $L(u) = K(u)u\mathbb{1}(u \geq 0)$ and $f(x) = \mu_A(x)$ in (5.3) to obtain

$$\mathbb{E}\left(K_h(X_i)\frac{T_iX_i}{h}\mathbb{E}(A \mid X_i)\right) = K_+^{(1)}\mu_{A+}f_X(0) + hK_+^{(2)}[\mu_A f_X]'_+ + \frac{h^2}{2}K_+^{(3)}[\mu_A f_X]''_+ + O(h^3).$$

Inserting that into (5.4) delivers

$$\begin{aligned}&\mathbb{E}(K_h(X_i)V_iA) \\ &= \underbrace{\begin{pmatrix}K_+^{(0)}\\K_+^{(0)}\\K_+^{(1)}\\K_+^{(1)}\end{pmatrix}\mu_{A+}f_X(0) + \begin{pmatrix}K_-^{(0)}\\0\\K_-^{(1)}\\0\end{pmatrix}\mu_{A-}f_X(0)}_{=:S_1} + \underbrace{h\left[\begin{pmatrix}K_+^{(1)}\\K_+^{(1)}\\K_+^{(2)}\\K_+^{(2)}\end{pmatrix}[\mu_A f_X]'_+ + \begin{pmatrix}K_-^{(1)}\\0\\K_-^{(2)}\\0\end{pmatrix}[\mu_A f_X]'_-\right]}_{=:S_2} \\ &\quad + \underbrace{\frac{h^2}{2}\left[\begin{pmatrix}K_+^{(2)}\\K_+^{(2)}\\K_+^{(3)}\\K_+^{(3)}\end{pmatrix}[\mu_A f_X]''_+ + \begin{pmatrix}K_-^{(2)}\\0\\K_-^{(3)}\\0\end{pmatrix}[\mu_A f_X]''_-\right]}_{=:S_3} + O(h^3).\end{aligned}$$

We consider all three summands separately. First remember the above properties of the

kernel and the fact that $\mu_{A+} = \tau_A + \mu_{A-}$. Then

$$\kappa(K)^{-1}S_1 = f_X(0)\kappa(K)^{-1}\left[\begin{pmatrix} K_+^{(0)} \\ K_+^{(0)} \\ K_+^{(1)} \\ K_+^{(1)} \end{pmatrix}(\tau_A + \mu_{A-}) + \begin{pmatrix} K_-^{(0)} \\ 0 \\ K_-^{(1)} \\ 0 \end{pmatrix}\mu_{A-}\right]$$

$$= f_X(0)\kappa(K)^{-1}\left[\tau_A \underbrace{\begin{pmatrix} K_+^{(0)} \\ K_+^{(0)} \\ K_+^{(1)} \\ K_+^{(1)} \end{pmatrix}}_{=:C_2} + \mu_{A-}\underbrace{\begin{pmatrix} 1 \\ K_+^{(0)} \\ 0 \\ K_+^{(1)} \end{pmatrix}}_{=:C_1}\right] = f_X(0)\begin{pmatrix} \mu_{A-} \\ \tau_A \\ 0 \\ 0 \end{pmatrix}$$

whereby the last equality holds since C_1 is the first column and C_2 the second column of $\kappa(K)$. Secondly, we obtain for the second summand

$$\kappa(K)^{-1}S_2 = h\kappa(K)^{-1}\left[\underbrace{\begin{pmatrix} K_+^{(1)} \\ K_+^{(1)} \\ K_+^{(2)} \\ K_+^{(2)} \end{pmatrix}}_{=:C_4}[\mu_A f_X]'_+ + \left(\underbrace{\begin{pmatrix} 0 \\ K_+^{(1)} \\ K_+^{(2)} \\ K_+^{(2)} \end{pmatrix}}_{=:C_3} - \underbrace{\begin{pmatrix} K_+^{(1)} \\ K_+^{(1)} \\ K_+^{(2)} \\ K_+^{(2)} \end{pmatrix}}_{=C_4}\right)[\mu_A f_X]'_-\right]$$

$$= h\begin{pmatrix} 0 \\ 0 \\ [\mu_A f_X]'_- \\ [\mu_A f_X]'_+ - [\mu_A f_X]'_- \end{pmatrix}$$

because C_3 and C_4 are the third and fourth column of $\kappa(K)$ respectively. Lastly, the third summand gives

$$\kappa(K)^{-1}S_3 = h^2 B(K, A) + O(h^3).$$

By taking everything together, we obtain

$$\kappa(K)^{-1}\mathbb{E}(K_h(X_i)V_i A) = \begin{pmatrix} f_X(0)\mu_{A-} \\ f_X(0)\tau_A \\ h[\mu_A f_X]'_- \\ h([\mu_A f_X]'_+ - [\mu_A f_X]'_-) \end{pmatrix} + h^2 B(K, A) + O(h^3)$$

which is the assertion of the Lemma. □

Lemma 5.3. *Let K be a kernel with $K^{(4)} < \infty$. Then,*

$$\kappa(K)^{-1} = \frac{1}{\left(K_+^{(1)}\right)^2 - \frac{1}{2}K_+^{(2)}} \begin{pmatrix} -K_+^{(2)} & K_+^{(2)} & -K_+^{(1)} & K_+^{(1)} \\ K_+^{(2)} & -2K_+^{(2)} & K_+^{(1)} & 0 \\ -K_+^{(1)} & K_+^{(1)} & -\frac{1}{2} & \frac{1}{2} \\ K_+^{(1)} & 0 & \frac{1}{2} & -1 \end{pmatrix}.$$

Furthermore define

$$a_1 = \frac{2\left(K_+^{(2)}\right)^2 - 2K_+^{(1)}K_+^{(3)}}{K_+^{(2)} - 2\left(K_+^{(1)}\right)^2}, \quad a_2 = \frac{K_+^{(3)} - 2K_+^{(1)}K_+^{(2)}}{K_+^{(2)} - 2\left(K_+^{(1)}\right)^2},$$
$$b_1 = \frac{2K_+^{(2)}K_+^{(3)} - 2K_+^{(1)}K_+^{(4)}}{K_+^{(2)} - 2\left(K_+^{(1)}\right)^2}, \quad b_2 = \frac{K_+^{(4)} - 2K_+^{(1)}K_+^{(3)}}{K_+^{(2)} - 2\left(K_+^{(1)}\right)^2},$$

then

$$\kappa(K)^{-1} \begin{pmatrix} 0 & K_+^{(1)} & 2K_+^{(2)} & K_+^{(2)} \\ K_+^{(1)} & K_+^{(1)} & K_+^{(2)} & K_+^{(2)} \\ 2K_+^{(2)} & K_+^{(2)} & 0 & K_+^{(3)} \\ K_+^{(2)} & K_+^{(2)} & K_+^{(3)} & K_+^{(3)} \end{pmatrix} = \begin{pmatrix} 0 & 0 & a_1 & 0 \\ 0 & 0 & 0 & a_1 \\ 1 & 0 & -a_2 & 0 \\ 0 & 1 & 2a_2 & a_2 \end{pmatrix}$$

and

$$\kappa(K)^{-1} \begin{pmatrix} 2K_+^{(2)} & K_+^{(2)} & 0 & K_+^{(3)} \\ K_+^{(2)} & K_+^{(2)} & K_+^{(3)} & K_+^{(3)} \\ 0 & K_+^{(3)} & 2K_+^{(4)} & K_+^{(4)} \\ K_+^{(3)} & K_+^{(3)} & K_+^{(4)} & K_+^{(4)} \end{pmatrix} = \begin{pmatrix} a_1 & 0 & -b_1 & 0 \\ 0 & a_1 & 2b_1 & b_1 \\ -a_2 & 0 & b_2 & 0 \\ 2a_2 & a_2 & 0 & b_2 \end{pmatrix}.$$

Proof. Jensen's inequality, stated in Proposition 1, implies $\left(K_+^{(1)}\right)^2 - \frac{1}{2}K_+^{(2)} \neq 0$ as shown in the proof of Lemma 5.1. The rest of the statement is just a direct calculation which we want to omit here. □

Lemma 5.4. *Let f_X be twice continuously differentiable in a neighborhood around zero and K be a kernel such that $K^{(4)} < \infty$. For $h \to 0$ and $nh \to \infty$ as $n \to \infty$, we have*

$$\begin{aligned} &\kappa(K)^{-1}\mathbb{E}(K_h(X_i)V_iV_i^\top) \\ &= f_X(0)I + f_X'(0)h \begin{pmatrix} 0 & 0 & a_1 & 0 \\ 0 & 0 & 0 & a_1 \\ 1 & 0 & -a_2 & 0 \\ 0 & 1 & 2a_2 & a_2 \end{pmatrix} + h^2\frac{f_X''(0)}{2} \begin{pmatrix} a_1 & 0 & -b_1 & 0 \\ 0 & a_1 & 2b_1 & b_1 \\ -a_2 & 0 & b_2 & 0 \\ 2a_2 & a_2 & 0 & b_2 \end{pmatrix} + o(h^2) \end{aligned}$$

whereby a_1, a_2, b_1, b_2 are defined as in Lemma 5.3.

Proof. We have

$$\begin{aligned}
&\mathbb{E}(K_h(X_i)V_iV_i^\top)\\
&= \mathbb{E}\left(K_h(X_i)\begin{pmatrix}1\\ T_i\\ \frac{X_i}{h}\\ \frac{T_iX_i}{h}\end{pmatrix}\begin{pmatrix}1 & T_i & \frac{X_i}{h} & \frac{T_iX_i}{h}\end{pmatrix}\right)\\
&= \begin{pmatrix}
\mathbb{E}(K_h(X_i)) & \mathbb{E}(K_h(X_i)T_i) & \mathbb{E}(K_h(X_i)\frac{X_i}{h}) & \mathbb{E}(K_h(X_i)\frac{T_iX_i}{h})\\
\mathbb{E}(K_h(X_i)T_i) & \mathbb{E}(K_h(X_i)T_i) & \mathbb{E}(K_h(X_i)\frac{T_iX_i}{h}) & \mathbb{E}(K_h(X_i)\frac{T_iX_i}{h})\\
\mathbb{E}(K_h(X_i)\frac{X_i}{h}) & \mathbb{E}(K_h(X_i)\frac{T_iX_i}{h}) & \mathbb{E}(K_h(X_i)\frac{X_i^2}{h^2}) & \mathbb{E}(K_h(X_i)\frac{T_iX_i^2}{h^2})\\
\mathbb{E}(K_h(X_i)\frac{T_iX_i}{h}) & \mathbb{E}(K_h(X_i)\frac{T_iX_i}{h}) & \mathbb{E}(K_h(X_i)\frac{T_iX_i^2}{h^2}) & \mathbb{E}(K_h(X_i)\frac{T_iX_i^2}{h^2})
\end{pmatrix}.
\end{aligned} \tag{5.5}$$

To calculate these means we use Taylor expansion as stated in (5.2) in combination with the kernel properties in (5.1), whereby we want to omit the detailed calculations here. In fact, this results in

- $\mathbb{E}(K_h(X_i)) = f_X(0) + h^2 f_X''(0)K_+^{(2)} + o(h^2)$ by setting $L = K$ and $f \equiv 1$ in (5.2)
- $\mathbb{E}(K_h(X_i)T_i) = K_+^{(0)} f_X(0) + h f_X'(0)K_+^{(1)} + \frac{h^2}{2} f_X''(0)K_+^{(2)} + o(h^2)$ by setting $L(u) = K(u)\mathbb{1}(u \geq 0)$ and $f \equiv 1$ in (5.2)
- $\mathbb{E}(K_h(X_i)\frac{X_i}{h}) = 2h f_X'(0)K_+^{(2)} + o(h^2)$ by setting $L(u) = K(u)u$ and $f \equiv 1$ in (5.2)
- $\mathbb{E}(K_h(X_i)\frac{T_iX_i}{h}) = f_X(0)K_+^{(1)} + h f_X'(0)K_+^{(2)} + \frac{h^2}{2} f_X''(0)K_+^{(3)} + o(h^2)$ by setting $L(u) = K(u)u\mathbb{1}(u \geq 0)$ and $f \equiv 1$ in (5.2)
- $\mathbb{E}(K_h(X_i)\frac{X_i^2}{h^2}) = 2f_X(0)K_+^{(2)} + h^2 f_X''(0)K_+^{(4)} + o(h^2)$ by setting $L(u) = K(u)u^2$ and $f \equiv 1$ in (5.2)
- $\mathbb{E}(K_h(X_i)\frac{T_iX_i^2}{h^2}) = f_X(0)K_+^{(2)} + h f_X'(0)K_+^{(3)} + \frac{h^2}{2} f_X''(0)K_+^{(4)} + o(h^2)$ by setting $L(u) = K(u)u^2\mathbb{1}(u \geq 0)$ and $f \equiv 1$ in (5.2).

Inserting that into equation (5.5) gives

$$\begin{aligned}
&\kappa(K)^{-1}\mathbb{E}(K_h(X_i)V_iV_i^\top)\\
&= f_X(0)I + f_X'(0)h\kappa(K)^{-1}\begin{pmatrix}
0 & K_+^{(1)} & 2K_+^{(2)} & K_+^{(2)}\\
K_+^{(1)} & K_+^{(1)} & K_+^{(2)} & K_+^{(2)}\\
2K_+^{(2)} & K_+^{(2)} & 0 & K_+^{(3)}\\
K_+^{(2)} & K_+^{(2)} & K_+^{(3)} & K_+^{(3)}
\end{pmatrix}\\
&\quad + \frac{h^2}{2} f_X''(0)\kappa(K)^{-1}\begin{pmatrix}
2K_+^{(2)} & K_+^{(2)} & 0 & K_+^{(3)}\\
K_+^{(2)} & K_+^{(2)} & K_+^{(3)} & K_+^{(3)}\\
0 & K_+^{(3)} & 2K_+^{(4)} & K_+^{(4)}\\
K_+^{(3)} & K_+^{(3)} & K_+^{(4)} & K_+^{(4)}
\end{pmatrix} + o(h^2).
\end{aligned}$$

We can finish the proof by using Lemma 5.3 now. □

From now on $\|\cdot\|_2$ applied on a vector denotes the Euclidean norm. When it is applied on a matrix, we mean the spectral norm, i.e. the operator norm induced by the Euclidean norm.

We will also have to deal with inverses of means in the subsequent content. In case we write down such an inverse in the assertion of one of the following lemmas, we always implicitly assume its existence in order to keep the statements more concise and to focus on the gist.

Lemma 5.5. *Suppose*

$$\left\|\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\right\|_2 = O(1)$$

and assume that f_X is continuous in a neighborhood around zero and μ_{ZY} can be extended continuously to zero from the left and the right. Then

$$\left\|\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\mathbb{E}(K_h(X_i)Z_iY_i)\right\|_2 = O(1).$$

Proof. By using a well-known property of the operator norm, we obtain

$$\left\|\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\mathbb{E}(K_h(X_i)Z_iY_i)\right\|_2 \leq \left\|\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\right\|_2 \left\|\mathbb{E}(K_h(X_i)Z_iY_i)\right\|_2.$$

The first factor is $O(1)$ by assumption. We can find an upper bound for the second factor by Taylor expansion:

$$\mathbb{E}(K_h(X_i)Z_iY_i) = \mathbb{E}(K_h(X_i)\mathbb{E}(Z_iY_i \mid X_i)) = \frac{1}{2}f_X(0)\mu_{ZY-} + \frac{1}{2}f_X(0)\mu_{ZY+} + o(1) = O(1).$$

Note that μ_{ZY-} and μ_{ZY+} exist by assumption. This shows the assertion of the lemma. □

Lemma 5.6. *Let K be a kernel with $K^{(4)} < \infty$, f_X be three times continuously differentiable in a neighborhood around zero with $f_X(0) > 0$ and $nh \to \infty, h \to 0$ as $n \to \infty$. Furthermore, suppose that μ_Z is continuous and one-sided differentiable at zero up to order three whereby the derivatives extend continuously to zero with $\lim_{x \nearrow 0} \mu_Z'(x) = \lim_{x \searrow 0} \mu_Z'(x) = 0$ and $\mu_Z(0) = 0$. Also, let*

$$\left\|\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\right\|_2 = O(1). \tag{5.6}$$

Then,

$$\begin{aligned}&\left[\left(I - \mathbb{E}(K_h(X_i)V_iV_i^\top)^{-1}\mathbb{E}(K_h(X_i)V_iZ_i^\top)\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\mathbb{E}(K_h(X_i)Z_iV_i^\top)\right)^{-1}\right]_2 \\ &= \begin{pmatrix}0 & 1 & 0 & 0\end{pmatrix} + O(h^4)\end{aligned}$$

Proof. Let $k \in \{1, ..., p\}$ and define the matrix $\gamma_n \in \mathbb{R}^{p \times 4}$ by its rows via

$$[\gamma_n]_{k,\cdot} := \mathbb{E}\left(K_h(X_i)V_i Z_i^{(k)}\right)^\top = \mathbb{E}\left(K_h(X_i)V_i \mathbb{E}\left(Z_i^{(k)} \mid X_i\right)\right)^\top$$
$$= \begin{pmatrix} \mathbb{E}\left(K_h(X_i)\mathbb{E}\left(Z_i^{(k)} \mid X_i\right)\right) \\ \mathbb{E}\left(K_h(X_i)T_i\mathbb{E}\left(Z_i^{(k)} \mid X_i\right)\right) \\ \mathbb{E}\left(K_h(X_i)\frac{X_i}{h}\mathbb{E}\left(Z_i^{(k)} \mid X_i\right)\right) \\ \mathbb{E}\left(K_h(X_i)\frac{T_iX_i}{h}\mathbb{E}\left(Z_i^{(k)} \mid X_i\right)\right) \end{pmatrix}^\top$$

We can examine each component by conducting a Taylor expansion as in (5.3) by setting

- $f(x) = \mu_{Z^{(k)}}(x), L(u) = K(u)$ for the first component,
- $f(x) = \mu_{Z^{(k)}}(x), L(u) = K(u)\mathbb{1}(u \geq 0)$ for the second component,
- $f(x) = \mu_{Z^{(k)}}(x), L(u) = K(u)u$ for the third component and
- $f(x) = \mu_{Z^{(k)}}(x), L(u) = K(u)u\mathbb{1}(u \geq 0)$ for the fourth component.

In addition, we also use the properties in (5.1) as well as the assumption that $\mu_Z(0) = 0$ and $\mu'_{Z+} = \mu'_{Z-} = 0$. Then, this results in

$$\begin{aligned}
[\gamma_n]_{k,\cdot}^\top &= \begin{pmatrix} K_+^{(0)} \\ K_+^{(0)} \\ K_+^{(1)} \\ K_+^{(1)} \end{pmatrix} \mu_{Z^{(k)}+} f_X(0) + \begin{pmatrix} K_-^{(0)} \\ 0 \\ K_-^{(1)} \\ 0 \end{pmatrix} \mu_{Z^{(k)}-} f_X(0) \\
&\quad + h\left[\begin{pmatrix} K_+^{(1)} \\ K_+^{(1)} \\ K_+^{(2)} \\ K_+^{(2)} \end{pmatrix} [\mu_{Z^{(k)}} f_X]'_+ + \begin{pmatrix} K_-^{(1)} \\ 0 \\ K_-^{(2)} \\ 0 \end{pmatrix} [\mu_{Z^{(k)}} f_X]'_-\right] \\
&\quad + \frac{h^2}{2}\left[\begin{pmatrix} K_+^{(2)} \\ K_+^{(2)} \\ K_+^{(3)} \\ K_+^{(3)} \end{pmatrix} [\mu_{Z^{(k)}} f_X]''_+ + \begin{pmatrix} K_-^{(2)} \\ 0 \\ K_-^{(3)} \\ 0 \end{pmatrix} [\mu_{Z^{(k)}} f_X]''_-\right] + O(h^3) \\
&= \frac{h^2}{2}\left[\begin{pmatrix} K_+^{(2)} \\ K_+^{(2)} \\ K_+^{(3)} \\ K_+^{(3)} \end{pmatrix} [\mu_{Z^{(k)}} f_X]''_+ + \begin{pmatrix} K_-^{(2)} \\ 0 \\ K_-^{(3)} \\ 0 \end{pmatrix} [\mu_{Z^{(k)}} f_X]''_-\right] + O(h^3) \\
&= \frac{h^2}{2}\begin{pmatrix} [\mu_{Z^{(k)}} f_X]''_+ K_+^{(2)} + [\mu_{Z^{(k)}} f_X]''_- K_-^{(2)} \\ [\mu_{Z^{(k)}} f_X]''_+ K_+^{(2)} \\ [\mu_{Z^{(k)}} f_X]''_+ K_+^{(3)} + [\mu_{Z^{(k)}} f_X]''_- K_-^{(3)} \\ [\mu_{Z^{(k)}} f_X]''_+ K_+^{(3)} \end{pmatrix} + O(h^3)
\end{aligned}$$

$$= \frac{h^2}{2}\left(\begin{pmatrix}[\mu_{Z^{(k)}}f_X]''_- & [\mu_{Z^{(k)}}f_X]''_+ - [\mu_{Z^{(k)}}f_X]''_-\end{pmatrix}\begin{pmatrix}K^{(2)} & K^{(2)}_+ & 0 & K^{(3)}_+\\ K^{(2)}_+ & K^{(2)}_+ & K^{(3)}_+ & K^{(3)}_+\end{pmatrix}\right)^\top + O(h^3).$$

Thus, we know that

$$\|\gamma_n\|_2 = O(h^2) + O(h^3) = O(h^2).$$

This in turn gives us

$$\begin{aligned}&\left\|\mathbb{E}(K_h(X_i)V_iZ_i^\top)\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\mathbb{E}(K_h(X_i)Z_iV_i^\top)\right\|_2\\ =\ &\left\|\gamma_n^\top\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\gamma_n\right\|_2 \le \underbrace{\left\|\gamma_n^\top\right\|_2}_{=O(h^2)}\underbrace{\left\|\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\right\|_2}_{=O(1)}\underbrace{\|\gamma_n\|_2}_{=O(h^2)} = O(h^4)\end{aligned} \tag{5.7}$$

whereby we used the submultiplicativity of the spectral norm. Also, we know by Lemma 5.4 that

$$\mathbb{E}(K_h(X_i)V_iV_i^\top)^{-1} = [f_X(0)\kappa(K) + O(h)]^{-1} = \frac{1}{f_X(0)}\kappa(K)^{-1} + O(h) = O(1). \tag{5.8}$$

Here we used that for $\epsilon_h = O(h)$ and A invertible and independent of h,

$$\begin{aligned}A^{-1} - (A+\epsilon_h)^{-1} &= A^{-1}(I - A(A+\epsilon_h)^{-1})\\ &= A^{-1}(A+\epsilon_h - A)(A+\epsilon_h)^{-1}\\ &= A^{-1}\underbrace{\epsilon_h}_{=O(h)}\underbrace{(A+\epsilon_h)^{-1}}_{=O(1)}\\ &= O(h)\end{aligned} \tag{5.9}$$

whereby $|\epsilon_h| \le Ch \xrightarrow{n\to\infty} 0$ and the continuity of the inverse implies $(A+\epsilon_h)^{-1} = O(1)$.

Taking the approximations (5.7) and (5.8) together, we obtain

$$I - \underbrace{\mathbb{E}(K_h(X_i)V_iV_i^\top)^{-1}}_{=O(1)}\underbrace{\gamma_n^\top\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\gamma_n}_{=O(h^4)} = I + O(h^4)$$

By using the same argument as in (5.9), we finally obtain the assertion of the lemma, i.e.

$$\left[\left(I - \mathbb{E}(K_h(X_i)V_iV_i^\top)^{-1}\gamma_n^\top\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\gamma_n\right)^{-1}\right]_2 = \begin{pmatrix}0 & 1 & 0 & 0\end{pmatrix} + O(h^4).$$

□

Now, we have laid the foundation for being able to prove the final lemma of this section, which provides us with a representation of the bias.

Lemma 5.7. *Let all assumptions of Lemma 5.5 and Lemma* 5.6 *hold. Furthermore, suppose that μ_Y is three times one-sided differentiable in zero, i.e. all the derivatives*

up to order three exist on $(-\infty, 0)$ *and* $(0, \infty)$*, and* μ_Y *as well as the derivatives extend continuously to zero from the left and the right side. Furthermore, define*

$$\mathcal{B}_n := \frac{1}{2}\frac{2(K_+^{(2)})^2 - 2K_+^{(1)}K_+^{(3)}}{K_+^{(2)} - 2(K_+^{(1)})^2}\Bigg(\mu''_{Y+} - \mu''_{Y-} - \sum_{k=1}^{p}\left(\mu''_{Z^{(k)}+} - \mu''_{Z^{(k)}-}\right)\left[\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\mathbb{E}(K_h(X_i)Z_iY_i)\right]_k\Bigg) + o(1).$$

Then,

$$\theta_0^{(2)} = \tau_Y + h^2\mathcal{B}_n.$$

Proof. Choose a_1, a_2, b_1, b_2 like in Lemma 5.3 and set

$$\kappa_{h,b}(K) = \left[\left(I + h\frac{f'_X(0)}{f_X(0)}\begin{pmatrix}0 & 0 & a_1 & 0\\ 0 & 0 & 0 & a_1\\ 1 & 0 & -a_2 & 0\\ 0 & 1 & 2a_2 & a_2\end{pmatrix} + h^2\frac{f''_X(0)}{2f_X(0)}\begin{pmatrix}a_1 & 0 & -b_1 & 0\\ 0 & a_1 & 2b_1 & b_1\\ -a_2 & 0 & b_2 & 0\\ 2a_2 & a_2 & 0 & b_2\end{pmatrix}\right)^{-1}\right]_2 - \begin{pmatrix}0 & 1 & 0 & 0\end{pmatrix}.$$

We know by definition that

$$(\theta_0(h), \gamma_0(h)) = \underset{(\theta,\gamma)\in\mathbb{R}^{p+4}}{\operatorname{argmin}}\ \mathbb{E}\left(K_h(X_i)(Y_i - \theta^\top V_i - \gamma^\top Z_i)^2\right).$$

By using the least squares algebra, we obtain

$$\begin{aligned}\begin{pmatrix}\theta_0(h)\\ \gamma_0(h)\end{pmatrix} &= \mathbb{E}\left(K_h(X_i)\begin{pmatrix}V_i\\ Z_i\end{pmatrix}\begin{pmatrix}V_i\\ Z_i\end{pmatrix}^\top\right)^{-1}\mathbb{E}\left(K_h(X_i)\begin{pmatrix}V_i\\ Z_i\end{pmatrix}Y_i\right)\\ &= \mathbb{E}\left(K_h(X_i)\begin{pmatrix}V_iV_i^\top & V_iZ_i^\top\\ Z_iV_i^\top & Z_iZ_i^\top\end{pmatrix}\right)^{-1}\mathbb{E}\left(K_h(X_i)\begin{pmatrix}V_i\\ Z_i\end{pmatrix}Y_i\right)\\ &= \begin{pmatrix}\mathbb{E}(K_h(X_i)V_iV_i^\top) & \mathbb{E}(K_h(X_i)V_iZ_i^\top)\\ \mathbb{E}(K_h(X_i)Z_iV_i^\top) & \mathbb{E}(K_h(X_i)Z_iZ_i^\top)\end{pmatrix}^{-1}\begin{pmatrix}\mathbb{E}(K_h(X_i)V_iY_i)\\ \mathbb{E}(K_h(X_i)Z_iY_i)\end{pmatrix}\end{aligned}$$

We calculate the inverse of the matrix by using the well-known formula for inverses of block matrices. Since our goal is to find a formula for $\theta_0^{(2)}$, we only need to calculate the upper blocks of the inverse.

$$\begin{aligned}&\begin{pmatrix}\mathbb{E}(K_h(X_i)V_iV_i^\top) & \mathbb{E}(K_h(X_i)V_iZ_i^\top)\\ \mathbb{E}(K_h(X_i)Z_iV_i^\top) & \mathbb{E}(K_h(X_i)Z_iZ_i^\top)\end{pmatrix}^{-1}\\ &= \begin{pmatrix}R & -R\mathbb{E}(K_h(X_i)V_iZ_i^\top)\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\\ * & *\end{pmatrix}\end{aligned}$$

whereby $R := (\mathbb{E}(K_h(X_i)V_iV_i^\top) - \mathbb{E}(K_h(X_i)V_iZ_i^\top)\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\mathbb{E}(K_h(X_i)Z_iV_i^\top))^{-1}$.

Denote the second row of R by R_2. Now, the second entry of θ_0 can be represented as following:

$$
\begin{aligned}
&\theta_0^{(2)} \\
&= R_2 \left(\mathbb{E}(K_h(X_i)V_iY_i) - \mathbb{E}\left(K_h(X_i)V_iZ_i^\top\right) \mathbb{E}\left(K_h(X_i)Z_iZ_i^\top\right)^{-1} \mathbb{E}(K_h(X_i)Z_iY_i) \right) \\
&= \left[\left(\mathbb{E}\left(K_h(X_i)V_iV_i^\top\right) - \mathbb{E}\left(K_h(X_i)V_iZ_i^\top\right) \mathbb{E}\left(K_h(X_i)Z_iZ_i^\top\right)^{-1} \mathbb{E}\left(K_h(X_i)Z_iV_i^\top\right) \right)^{-1} \right]_2 \\
&\quad \left(\mathbb{E}(K_h(X_i)V_iY_i) - \mathbb{E}\left(K_h(X_i)V_iZ_i^\top\right) \mathbb{E}\left(K_h(X_i)Z_iZ_i^\top\right)^{-1} \mathbb{E}(K_h(X_i)Z_iY_i) \right) \\
&= \Bigg[\Bigg(\mathbb{E}\left(K_h(X_i)V_iV_i^\top\right) \Big(I - \mathbb{E}\left(K_h(X_i)V_iV_i^\top\right)^{-1} \mathbb{E}\left(K_h(X_i)V_iZ_i^\top\right) \mathbb{E}\left(K_h(X_i)Z_iZ_i^\top\right)^{-1} \\
&\quad \mathbb{E}\left(K_h(X_i)Z_iV_i^\top\right) \Big) \Bigg)^{-1} \Bigg]_2 \qquad (5.10) \\
&\quad \left(\mathbb{E}(K_h(X_i)V_iY_i) - \mathbb{E}\left(K_h(X_i)V_iZ_i^\top\right) \mathbb{E}\left(K_h(X_i)Z_iZ_i^\top\right)^{-1} \mathbb{E}(K_h(X_i)Z_iY_i) \right) \\
&= \Bigg[\Big(I - \mathbb{E}\left(K_h(X_i)V_iV_i^\top\right)^{-1} \mathbb{E}\left(K_h(X_i)V_iZ_i^\top\right) \mathbb{E}\left(K_h(X_i)Z_iZ_i^\top\right)^{-1} \\
&\quad \mathbb{E}\left(K_h(X_i)Z_iV_i^\top\right) \Big)^{-1} \Bigg]_2 \left(\kappa(K)^{-1} \mathbb{E}\left(K_h(X_i)V_iV_i^\top\right) \right)^{-1} \\
&\quad \kappa(K)^{-1} \left(\mathbb{E}(K_h(X_i)V_iY_i) - \mathbb{E}\left(K_h(X_i)V_iZ_i^\top\right) \mathbb{E}\left(K_h(X_i)Z_iZ_i^\top\right)^{-1} \mathbb{E}(K_h(X_i)Z_iY_i) \right).
\end{aligned}
$$

By applying Lemma 5.6 and Lemma 5.4 we obtain that

$$
\begin{aligned}
&\Bigg[\Big(I - \mathbb{E}\left(K_h(X_i)V_iV_i^\top\right)^{-1} \mathbb{E}\left(K_h(X_i)V_iZ_i^\top\right) \mathbb{E}\left(K_h(X_i)Z_iZ_i^\top\right)^{-1} \\
&\quad \mathbb{E}\left(K_h(X_i)Z_iV_i^\top\right) \Big)^{-1} \Bigg]_2 \left(\kappa(K)^{-1} \mathbb{E}\left(K_h(X_i)V_iV_i^\top\right) \right)^{-1} \\
&= \left(\begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix} + O(h^4) \right) \\
&\quad \cdot \left(f_X(0) I + h f_X'(0) \begin{pmatrix} 0 & 0 & a_1 & 0 \\ 0 & 0 & 0 & a_1 \\ 1 & 0 & -a_2 & 0 \\ 0 & 1 & 2a_2 & a_2 \end{pmatrix} + h^2 \frac{f_X''(0)}{2} \begin{pmatrix} a_1 & 0 & -b_1 & 0 \\ 0 & a_1 & 2b_1 & b_1 \\ -a_2 & 0 & b_2 & 0 \\ 2a_2 & a_2 & 0 & b_2 \end{pmatrix} + o(h^2) \right)^{-1} \\
&\overset{(*)}{=} \frac{1}{f_X(0)} \left[\left(I + h \frac{f_X'(0)}{f_X(0)} \begin{pmatrix} 0 & 0 & a_1 & 0 \\ 0 & 0 & 0 & a_1 \\ 1 & 0 & -a_2 & 0 \\ 0 & 1 & 2a_2 & a_2 \end{pmatrix} + h^2 \frac{f_X''(0)}{2f_X(0)} \begin{pmatrix} a_1 & 0 & -b_1 & 0 \\ 0 & a_1 & 2b_1 & b_1 \\ -a_2 & 0 & b_2 & 0 \\ 2a_2 & a_2 & 0 & b_2 \end{pmatrix} \right)^{-1} \right]_2 \\
&\quad + o(h^2) \\
&= \frac{1}{f_X(0)} \left(\begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix} + \kappa_{h,b}(K) \right) + o(h^2).
\end{aligned}
$$

Note that the equality $(*)$ uses the fact that $(A_h+\epsilon_h)^{-1} = A_h^{-1} + o(h^2)$ if A_h is invertible, $A_h \to A$ for an invertible A, and $\epsilon_h = o(h^2)$ for $h \to 0$. This is due to

$$\begin{aligned} A_h^{-1} - (A_h+\epsilon_h)^{-1} &= A_h^{-1}(I - A_h(A_h+\epsilon_h)^{-1}) \\ &= A_h^{-1}(A_h+\epsilon_h - A_h)(A_h+\epsilon_h)^{-1} \\ &= \underbrace{A_h^{-1}}_{=O(1)} \underbrace{\epsilon_h}_{=o(h^2)} \underbrace{(A_h+\epsilon_h)^{-1}}_{=O(1)} = o(h^2) \end{aligned}$$

whereby $(A_h+\epsilon_h)^{-1} = O(1)$ follows from the fact that in particular $\epsilon_h \to 0$, $A_h \to A$ for $h \to 0$ and that the inverse function is continuous. Also note that the invertibility of $A_h + \epsilon_h$ is implied by the openness of the set of invertible matrices.

We set $A = Y_i$ in Lemma 5.2 and obtain

$$\kappa(K)^{-1}\mathbb{E}(K_h(X_i)V_iY_i) = \begin{pmatrix} f_X(0)\mu_{Y-} \\ f_X(0)\tau_Y \\ h(\mu_Y f_X)'_- \\ h((\mu_Y f_X)'_+ - (\mu_Y f_X)'_-) \end{pmatrix} + h^2 B(K,Y) + O(h^3).$$

Also, we can apply this lemma with $A = Z_i^{(k)}$ for $1 \le k \le p$:

$$\begin{aligned} &\kappa(K)^{-1}\mathbb{E}(K_h(X_i)V_iZ_i^\top) \\ &= \kappa(K)^{-1}\mathbb{E}\left(K_h(X_i)V_i\left(\sum_{k=1}^p Z_i^{(k)}e_k^\top\right)\right) \\ &= \sum_{k=1}^p \kappa(K)^{-1}\mathbb{E}\left(K_h(X_i)V_iZ_i^{(k)}\right)e_k^\top \\ &= \sum_{k=1}^p \left(\begin{pmatrix} f_X(0)\mu_{Z^{(k)}-} \\ f_X(0)\tau_{Z^{(k)}} \\ h(\mu_{Z^{(k)}}f_X)'_- \\ h((\mu_{Z^{(k)}}f_X)'_+ - (\mu_{Z^{(k)}}f_X)'_-) \end{pmatrix} + h^2 B\left(K, Z_i^{(k)}\right) + O(h^3)\right)e_k^\top \\ &= \sum_{k=1}^p \left(h^2 B\left(K, Z_i^{(k)}\right) + O(h^3)\right)e_k^\top \\ &= h^2 B + O(h^3) \end{aligned} \tag{5.11}$$

whereby B is defined as the matrix which has $B(K, Z_i^{(k)}), k = 1, ..., p$ as columns. Note that we used the assumptions $\mu_{Z^{(k)}}(0) = 0$ as well as $\lim_{x \searrow 0} \mu'_{Z^{(k)}}(x) = \lim_{x \nearrow 0} \mu'_{Z^{(k)}}(x) = 0$, which also implies $\tau_{Z^{(k)}} = 0$. Now, we can state that

$$\begin{aligned} &\kappa(K)^{-1}\mathbb{E}(K_h(X_i)V_iZ_i^\top)\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\mathbb{E}(K_h(X_i)Z_iY_i) \\ &= h^2 B\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\mathbb{E}(K_h(X_i)Z_iY_i) + O(h^3)\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\mathbb{E}(K_h(X_i)Z_iY_i). \end{aligned}$$

We want to find an upper bound for that. This can be done by noticing that

$$\left\|\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\mathbb{E}(K_h(X_i)Z_iY_i)\right\|_2 = O(1)$$

by Lemma 5.5. This leads to

$$\begin{aligned}&\left[\kappa(K)^{-1}\mathbb{E}(K_h(X_i)V_iZ_i^\top)\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\mathbb{E}(K_h(X_i)Z_iY_i)\right]_l\\ &\leq h^2\left\|B_{l\cdot}\right\|_2 O(1) + \left\|(O(h^3))_{l\cdot}\right\|_2 O(1) = O(h^2)+O(h^3) = O(h^2)\end{aligned}$$

for $l \in \{1,...,4\}$. Now we merge all of the statements above to obtain

$$\begin{aligned}&\theta_0^{(2)}\\
&= \left[\frac{1}{f_X(0)}\left(\begin{pmatrix}0&1&0&0\end{pmatrix}+\kappa_{h,b}(K)\right)+o(h^2)\right]\\
&\quad\cdot\kappa(K)^{-1}\left(\mathbb{E}(K_h(X_i)V_iY_i)-\mathbb{E}(K_h(X_i)V_iZ_i^\top)\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\mathbb{E}(K_h(X_i)Z_iY_i)\right)\\
&= \left[\frac{1}{f_X(0)}\left(\begin{pmatrix}0&1&0&0\end{pmatrix}+\kappa_{h,b}(K)\right)+o(h^2)\right]\\
&\quad\cdot\Bigg[\begin{pmatrix}f_X(0)\mu_{Y-}\\ f_X(0)\tau_Y\\ h(\mu_Y f_X)'_-\\ h((\mu_Y f_X)'_+-(\mu_Y f_X)'_-)\end{pmatrix}+h^2B(K,Y)+O(h^3)\\
&\qquad-h^2B\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\mathbb{E}(K_h(X_i)Z_iY_i)-\underbrace{O(h^3)\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\mathbb{E}(K_h(X_i)Z_iY_i)}_{=O(h^3)}\Bigg]\\
&= \tau_Y+\frac{h^2}{f_X(0)}[B(K,Y)]_2-\frac{h^2}{f_X(0)}B_{2\cdot}\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\mathbb{E}(K_h(X_i)Z_iY_i)\\
&\quad+\underbrace{[O(h^3)]_2}_{=O(h^3)}-\frac{1}{f_X(0)}\underbrace{\kappa_{h,b}(K)}_{=o(1)}\underbrace{(h^2B\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\mathbb{E}(K_h(X_i)Z_iY_i)+O(h^3))}_{=O(h^2)}\\
&\quad+\kappa_{h,b}(K)\begin{pmatrix}\mu_{Y-}\\ \tau_Y\\ \frac{h}{f_X(0)}(\mu_Y f_X)'_-\\ \frac{h}{f_X(0)}((\mu_Y f_X)'_+-(\mu_Y f_X)'_-)\end{pmatrix}+\underbrace{\frac{\kappa_{h,b}(K)}{f_X(0)}h^2B(K,Y)}_{=o(h^2)}+\underbrace{\frac{\kappa_{h,b}(K)}{f_X(0)}O(h^3)}_{=o(h^2)}\\
&\quad+\underbrace{o(h^2)\begin{pmatrix}f_X(0)\mu_{Y-}\\ f_X(0)\tau_Y\\ h(\mu_Y f_X)'_-\\ h((\mu_Y f_X)'_+-(\mu_Y f_X)'_-)\end{pmatrix}}_{=o(h^2)}+\underbrace{o(h^2)h^2B(K,Y)}_{=o(h^2)}+\underbrace{o(h^2)O(h^2)}_{=o(h^2)}\end{aligned}$$

$$= \tau_Y + \frac{h^2}{f_X(0)}[B(K,Y)]_2 - \frac{h^2}{f_X(0)} B_2 . \mathbb{E}(K_h(X_i) Z_i Z_i^\top)^{-1} \mathbb{E}(K_h(X_i) Z_i Y_i)$$
$$+ \kappa_{h,b}(K) \begin{pmatrix} \mu_{Y-} \\ \tau_Y \\ \frac{h}{f_X(0)}(\mu_Y f_X)'_- \\ \frac{h}{f_X(0)}((\mu_Y f_X)'_+ - (\mu_Y f_X)'_-) \end{pmatrix} + o(h^2),$$

whereby we used that $\kappa_{h,b}(K) = o(1)$ which is shown by taking the definition and letting $h \to 0$. The main part of the proof is done. The rest is conducted by just following some straightforward computations. As the next calculation is rather long and technical, we want to refer to using a CAS in order to obtain

$$\kappa_{h,b}(K) = \begin{pmatrix} o(h^2) & h^2 a_1 \left(\frac{f'_X(0)^2}{f_X(0)^2} - \frac{f''_X(0)}{2 f_X(0)} \right) + o(h^2) & O(h^2) & -h a_1 \frac{f'_X(0)}{f_X(0)} + O(h^2) \end{pmatrix}.$$

We now consider the result of Lemma 5.3 by examining the matrix multiplication column-wise, which gives us that

$$\kappa(K)^{-1} \begin{pmatrix} K_+^{(2)} \\ K_+^{(2)} \\ K_+^{(3)} \\ K_+^{(3)} \end{pmatrix} = \begin{pmatrix} 0 \\ a_1 \\ 0 \\ a_2 \end{pmatrix}$$

and

$$\kappa(K)^{-1} \begin{pmatrix} K_-^{(2)} \\ 0 \\ K_-^{(3)} \\ 0 \end{pmatrix} = \kappa(K)^{-1} \begin{pmatrix} K_+^{(2)} \\ 0 \\ -K_+^{(3)} \\ 0 \end{pmatrix} = \kappa(K)^{-1} \left[\begin{pmatrix} 2K_+^{(2)} \\ K_+^{(2)} \\ 0 \\ K_+^{(3)} \end{pmatrix} - \begin{pmatrix} K_+^{(2)} \\ K_+^{(2)} \\ K_+^{(3)} \\ K_+^{(3)} \end{pmatrix} \right]$$
$$= \begin{pmatrix} a_1 \\ 0 \\ -a_2 \\ 2a_2 \end{pmatrix} - \begin{pmatrix} 0 \\ a_1 \\ 0 \\ a_2 \end{pmatrix} = \begin{pmatrix} a_1 \\ -a_1 \\ -a_2 \\ a_2 \end{pmatrix}$$

whereby we used that $K_-^{(2)} = K_+^{(2)}$ and $K_-^{(3)} = -K_+^{(3)}$ since the kernel is symmetric. Replacing the above identities in the definition of $B(K,Y)$ delivers

$$[B(K,Y)]_2$$
$$= \frac{a_1}{2} \left(\mu''_{Y+} f_X(0) + 2\mu'_{Y+} f'_X(0) + \mu_{Y+} f''_X(0) - \mu''_{Y-} f_X(0) - 2\mu'_{Y-} f'_X(0) - \mu_{Y-} f''_X(0) \right).$$

Therefore

$$
\begin{aligned}
&\frac{h^2}{f_X(0)}[B(K,Y)]_2 + \kappa_{h,b}(K)\begin{pmatrix} \mu_{Y-} \\ \tau_Y \\ \frac{h}{f_X(0)}(\mu_Y f_X)'_- \\ \frac{h}{f_X(0)}((\mu_Y f_X)'_+ - (\mu_Y f_X)'_-) \end{pmatrix} \\
&= \frac{h^2}{f_X(0)}\frac{a_1}{2}\left(f_X(0)\left(\mu''_{Y+}-\mu''_{Y-}\right)+2f'_X(0)\left(\mu'_{Y+}-\mu'_{Y-}\right)+f''_X(0)\left(\mu_{Y+}-\mu_{Y-}\right)\right) \\
&\quad + \underbrace{o(h^2)\mu_{Y-}}_{=o(h^2)} + h^2 a_1\left(\frac{f'_X(0)^2}{f_X(0)^2}-\frac{f''_X(0)}{2f_X(0)}\right)\left(\mu_{Y+}-\mu_{Y-}\right) + \underbrace{o(h^2)\tau_Y}_{=o(h^2)} \\
&\quad + \underbrace{O(h^2)\frac{h}{f_X(0)}(\mu_Y f_X)'_-}_{=o(h^2)} - h a_1 \frac{f'_X(0)}{f_X(0)}\frac{h}{f_X(0)}\left((\mu_Y f_X)'_+ - (\mu_Y f_X)'_-\right) \\
&\quad + \underbrace{O(h^2)\frac{h}{f_X(0)}((\mu_Y f_X)'_+ - (\mu_Y f_X)'_-)}_{=o(h^2)} \\
&= h^2 a_1\Bigg(\frac{1}{2}\left(\mu''_{Y+}-\mu''_{Y-}\right)+\frac{f'_X(0)}{f_X(0)}\left(\mu'_{Y+}-\mu'_{Y-}\right)+\frac{1}{2}\frac{f''_X(0)}{f_X(0)}\left(\mu_{Y+}-\mu_{Y-}\right) \\
&\quad + \frac{f'_X(0)^2}{f_X(0)^2}\left(\mu_{Y+}-\mu_{Y-}\right)-\frac{1}{2}\frac{f''_X(0)}{f_X(0)}\left(\mu_{Y+}-\mu_{Y-}\right) \\
&\quad - \frac{f'_X(0)}{f_X(0)^2}\left(\mu'_{Y+}f_X(0)+\mu_{Y+}f'_X(0)-\mu'_{Y-}f_X(0)-\mu_{Y-}f'_X(0)\right)\Bigg)+o(h^2) \\
&= h^2\frac{a_1}{2}\left(\mu''_{Y+}-\mu''_{Y-}\right)+o(h^2).
\end{aligned}
$$

Using a representation of $B(K, Z_i^{(k)})$ in the above manner again, gives that

$$
\begin{aligned}
&\frac{h^2}{f_X(0)}B_2.\mathbb{E}\left(K_h(X_i)Z_iZ_i^\top\right)^{-1}\mathbb{E}(K_hZ_iY_i) \\
&= h^2\frac{a_1}{2}\sum_{k=1}^{p}\left(\mu''_{Z^{(k)}+}-\mu''_{Z^{(k)}-}\right)\left[\mathbb{E}\left(K_h(X_i)Z_iZ_i^\top\right)^{-1}\mathbb{E}(K_hZ_iY_i)\right]^{(k)}
\end{aligned}
$$

since $\mu_{Z+} = \mu_{Z-}$ and $\mu'_{Z+} = \mu'_{Z-}$. Combining the above two statements provides

$$
\begin{aligned}
\theta_0^{(2)} &= \tau_Y + h^2\frac{a_1}{2}\Bigg(\left(\mu''_{Y+}-\mu''_{Y-}\right) \\
&\qquad + \sum_{k=1}^{p}\left(\mu''_{Z^{(k)}+}-\mu''_{Z^{(k)}-}\right)\left[\mathbb{E}\left(K_h(X_i)Z_iZ_i^\top\right)^{-1}\mathbb{E}(K_hZ_iY_i)\right]^{(k)}\Bigg)+o(h^2) \\
&= \tau_Y + h^2\mathcal{B}_n,
\end{aligned}
$$

which proves the Lemma. □

6 Asymptotic Behavior

This chapter lays the foundation for the second main step of proving asymptotic normality of the covariate-adjusted average treatment effect estimator. In that step, this chapter will be used by applying it on an adjusted version of the covariates, namely $\tilde{Z}_i = Z_i - M_n^\top V_i$ for some M_n that will be defined later on. By doing that, some additional properties concerning the value of $\mu_{\tilde{Z}}$ and of the one-sided limits of $\mu'_{\tilde{Z}}$ in zero emerge, which are necessary to fulfill the assumptions of the assertions contained in this chapter. Again, some of the results are based on adjustments of ideas from [KR21] to the case of a fixed number of covariates.

Before we start with a straightforward statement on the convergence of the sample mean, please recall Definition 4.1 which defines a kernel to be non-negative, integrable, integrating to one and symmetric.

Lemma 6.1. *Let K be a kernel which is compactly supported on $[-1,1]$ and satisfies $(K^2)^{(0)} < \infty$. Let f_X be continuous in a neighborhood around zero and $h \to 0, nh \to \infty$ as $n \to \infty$. Moreover, let $A_i \in \mathbb{R}, i = 1, ..., n$ be random variables, which can also depend on n and h, such that $((X_i, A_i))_{i=1,...,n}$ is a family of n independent random variables and*

$$\sup_{n\in\mathbb{N}} \sup_{x\in[-h,h]} \mathbb{E}(A_i^2 \mid X_i = x) < \infty. \tag{6.1}$$

Then,

$$\frac{1}{n}\sum_{i=1}^{n} K_h(X_i)A_i = \mathbb{E}(K_h(X_i)A_i) + O_P\left(\frac{1}{\sqrt{nh}}\right).$$

Proof. We set

$$S_n = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i)A_i.$$

By the imposed assumptions, we know that the mean and variance of S_n are finite. Therefore, we can use Chebyshev's inequality in order to obtain

$$\begin{aligned}
&P(|S_n - \mathbb{E}(S_n)| > M) \\
&\leq \frac{1}{M^2}\mathbb{E}\left((S_n - \mathbb{E}(S_n))^2\right) \\
&= \frac{1}{M^2}\mathbb{E}\left(\left(\frac{1}{n}\sum_{i=1}^{n} K_h(X_i)A_i - \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} K_h(X_i)A_i\right)\right)^2\right) \\
&= \frac{1}{M^2}\frac{1}{n^2}\mathbb{E}\left(\left(\sum_{i=1}^{n}\left(K_h(X_i)A_i - \mathbb{E}(K_h(X_i)A_i)\right)\right)^2\right) \\
&= \frac{1}{M^2}\frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^{n}(K_h(X_i)A_i - \mathbb{E}(K_h(X_i)A_i))^2\right) + \sum_{i\neq j}\underbrace{\mathrm{Cov}(K_h(X_i)A_i, K_h(X_j)A_j)}_{=0}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{M^2}\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left((K_h(X_i)A_i - \mathbb{E}(K_h(X_i)A_i))^2\right)\\
&= \frac{1}{M^2}\frac{1}{n}\mathbb{E}\left((K_h(X_i)A_i)^2 - 2K_h(X_i)A_i\mathbb{E}(K_h(X_i)A_i) + \mathbb{E}(K_h(X_i)A_i)^2\right)\\
&= \frac{1}{M^2}\frac{1}{n}\mathbb{E}\left((K_h(X_i)A_i)^2 - K_h(X_i)A_i\mathbb{E}(K_h(X_i)A_i)\right)\\
&\leq \frac{1}{M^2}\frac{1}{n}\mathbb{E}\left((K_h(X_i)A_i)^2\right)\\
&= \frac{1}{M^2}\frac{1}{n}\mathbb{E}\left(\mathbb{E}\left((K_h(X_i)A_i)^2 \mid X_i\right)\right)\\
&= \frac{1}{M^2}\frac{1}{n}\mathbb{E}\left(K_h(X_i)^2\mathbb{E}\left(A_i^2 \mid X_i\right)\right)\\
&\leq \sup_{n\in\mathbb{N}}\sup_{x\in[-h,h]}\mathbb{E}\left(A_i^2 \mid X_i = x\right)\frac{1}{M^2}\frac{1}{n}\frac{1}{h^2}\int_{-h}^{h}K\left(\frac{x}{h}\right)^2 f_X(x)\,dx\\
&= \sup_{n\in\mathbb{N}}\sup_{x\in[-h,h]}\mathbb{E}\left(A_i^2 \mid X_i = x\right)\frac{1}{M^2}\frac{1}{nh}\int_{-1}^{1}K\left(y\right)^2 f_X(yh)\,dy = O\left(\frac{1}{nhM^2}\right),
\end{aligned}
$$

whereby $\int_{-1}^{1}K\left(y\right)^2 f_X(yh)\,dy = O(1)$ due to $(K^2)^{(0)} < \infty$, $h \to 0$ and the continuity of f_X in a neighborhood around zero. Overall, this leads to

$$P\left(|S_n - \mathbb{E}(S_n)| > M\right) \leq \frac{C}{nhM^2}$$

for some $C > 0$. Given an arbitrary $\epsilon > 0$, choose M large enough such that $\frac{C}{M^2} < \epsilon$. Then,

$$P(\sqrt{nh}|S_n - \mathbb{E}(S_n)| > M) = P\left(|S_n - \mathbb{E}(S_n)| > \frac{M}{\sqrt{nh}}\right) \leq \frac{C}{M^2} < \epsilon,$$

which means that $S_n - \mathbb{E}(S_n) = O_P\left(\frac{1}{\sqrt{nh}}\right)$. □

As we need a matrix notation for our population data in order to write the statements of the subsequent lemmas in a more compact and concise way, the following definition will be highly beneficial.

Definition 6.1. *Let K be a kernel function, then we set*

$$\mathbf{K}_h = \operatorname{diag}\left(\frac{1}{h}K\left(\frac{X_1}{h}\right), ..., \frac{1}{h}K\left(\frac{X_n}{h}\right)\right)$$

and

$$\mathbf{K}_h^{\frac{1}{2}} = \operatorname{diag}\left(\sqrt{\frac{1}{h}K\left(\frac{X_1}{h}\right)}, ..., \sqrt{\frac{1}{h}K\left(\frac{X_n}{h}\right)}\right).$$

Furthermore, we define

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbb{R}^n, \quad \mathbf{Z} = \begin{pmatrix} Z_1^\top \\ \vdots \\ Z_n^\top \end{pmatrix} \in \mathbb{R}^{n\times p}, \quad \mathbf{r}(h) = \begin{pmatrix} r_1(h) \\ \vdots \\ r_n(h) \end{pmatrix} \in \mathbb{R}^n$$

and

$$\mathbf{V} = \begin{pmatrix} 1 & T_1 & X_1/h & T_1X_1/h \\ \vdots & \vdots & \vdots & \vdots \\ 1 & T_n & X_n/h & T_nX_n/h \end{pmatrix} \in \mathbb{R}^{n\times 4}.$$

This notation allows us to concisely make statements about the asymptotic behaviour of our population data within the next two lemmas.

Lemma 6.2. *Let K be a kernel which is compactly supported on $[-1,1]$ and satisfies $(K^2)^{(0)} < \infty$. Let f_X be continuous in a neighborhood around zero. Furthermore, let $h \to 0, nh \to \infty$ as $n \to \infty$ and*

$$\sup_{n\in\mathbb{N}} \sup_{x\in[-h,h]} \mathbb{E}(r_i(h)^2 \mid X_i = x) < \infty. \tag{6.2}$$

Then,

$$\frac{1}{n}\mathbf{V}^\top \mathbf{K}_h \mathbf{r}(h) = O_P\left(\frac{1}{\sqrt{nh}}\right).$$

Proof. We want to prove this row-wise. Let $a \in \{1, ..., 4\}$, then

$$\begin{aligned}\mathbb{E}\left(\frac{1}{h}K\left(\frac{X_i}{h}\right)^2 V_{i,a}^2 r_i(h)^2\right) &= \mathbb{E}\left(\mathbb{E}\left(\frac{1}{h}K\left(\frac{X_i}{h}\right)^2 V_{i,a}^2 r_i(h)^2 \,\middle|\, X_i\right)\right) \\ &= \mathbb{E}\left(\frac{1}{h}K\left(\frac{X_i}{h}\right)^2 V_{i,a}^2 \mathbb{E}(r_i(h)^2 \mid X_i)\right) = O(1)\end{aligned}$$

due to assumption (6.2) and our assumptions imposed on f_X and K. Using Markov's inequality, we obtain for every $\epsilon > 0$ that

$$\begin{aligned}P\left(\left|\frac{1}{n}\mathbf{V}_{\cdot a}^\top \mathbf{K}_h \mathbf{r}(h)\right| > M\right) &\le \frac{1}{M^2}\mathbb{E}\left(\left(\frac{1}{n}\sum_{i=1}^n \frac{1}{h}K\left(\frac{X_i}{h}\right) V_{i,a} r_i(h)\right)^2\right) \\ &= \frac{1}{nhM^2}\mathbb{E}\left(\frac{1}{h}K\left(\frac{X_i}{h}\right)^2 V_{i,a}^2 r_i(h)^2\right) = O\left(\frac{1}{nhM^2}\right)\end{aligned}$$

leading to

$$P\left(\left|\frac{1}{n}\mathbf{V}_{\cdot a}^\top \mathbf{K}_h \mathbf{r}(h)\right| > M\right) \le \frac{C}{nhM^2}$$

for some constant $C > 0$. Now we can conclude with the same argument as in the proof of Lemma 6.1. □

Lemma 6.3. *Let all assumptions of Lemma* 5.6 *hold. Also, let K be a kernel which is compactly supported on $[-1,1]$ and satisfies $K^{(4)}, (K^2)^{(0)} < \infty$. Additionally, define*

$$M := I - \mathbf{K}_h^{\frac{1}{2}} \mathbf{Z} (\mathbf{Z}^\top \mathbf{K}_h \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{K}_h^{\frac{1}{2}}$$

and suppose that for all $k, l \in \{1, ..., p\}$

$$\begin{aligned} &\sup_{n \in \mathbb{N}} \sup_{x \in [-h,h]} \mathbb{E}\left(\left(Z_i^{(k)} r_i(h) \right)^2 \,\middle|\, X_i = x \right) < \infty, \\ &\sup_{n \in \mathbb{N}} \sup_{x \in [-h,h]} \mathbb{E}\left(\left(Z_i^{(k)} Z_i^{(l)} \right)^2 \,\middle|\, X_i = x \right) < \infty. \end{aligned} \tag{6.3}$$

Then

$$\frac{1}{n} \mathbf{V}^\top \mathbf{K}_h^{\frac{1}{2}} M \mathbf{K}_h^{\frac{1}{2}} \mathbf{V} = f_X(0) \kappa(K) + o_P(1)$$

and

$$\frac{1}{n} \mathbf{V}^\top \mathbf{K}_h^{\frac{1}{2}} M \mathbf{K}_h^{\frac{1}{2}} \mathbf{r}(h) = \frac{1}{n} \mathbf{V}^\top \mathbf{K}_h \mathbf{r}(h) + o_P\left(\frac{1}{\sqrt{nh}} \right).$$

Proof. The idea of the proof is to use Lemma 6.1 for multiple times. Indeed, we can do this by calculating

$$\begin{aligned} \frac{1}{n} \mathbf{V}^\top \mathbf{K}_h^{\frac{1}{2}} M \mathbf{K}_h^{\frac{1}{2}} \mathbf{V} &= \frac{1}{n} \mathbf{V}^\top \mathbf{K}_h^{\frac{1}{2}} \left(I - \mathbf{K}_h^{\frac{1}{2}} \mathbf{Z} (\mathbf{Z}^\top \mathbf{K}_h \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{K}_h^{\frac{1}{2}} \right) \mathbf{K}_h^{\frac{1}{2}} \mathbf{V} \\ &= \frac{1}{n} \mathbf{V}^\top \mathbf{K}_h \mathbf{V} - \frac{1}{n} \mathbf{V}^\top \mathbf{K}_h \mathbf{Z} (\mathbf{Z}^\top \mathbf{K}_h \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{K}_h \mathbf{V} \\ &= \underbrace{\frac{1}{n} \mathbf{V}^\top \mathbf{K}_h \mathbf{V}}_{(1)} - \underbrace{\left(\frac{1}{n} \mathbf{V}^\top \mathbf{K}_h \mathbf{Z} \right) \left(\frac{1}{n} \mathbf{Z}^\top \mathbf{K}_h \mathbf{Z} \right)^{-1} \left(\frac{1}{n} \mathbf{Z}^\top \mathbf{K}_h \mathbf{V} \right)}_{(2)}. \end{aligned}$$

Concerning term (2), we can now estimate each of the three factors individually. For the first factor we obtain

$$\frac{1}{n} \mathbf{V}^\top \mathbf{K}_h \mathbf{Z} = \frac{1}{n} \sum_{i=1}^n K_h(X_i) V_i Z_i^\top$$

leading to the fact that we can consider that entry-wise and respectively set A_i of Lemma 6.1 to every entry of $V_i Z_i^\top$. The necessary assumption (6.1) for that is implied by (6.3) as we have for $m \in \{1, ..., 4\}$, $k \in \{1, ..., p\}$ and $u(x) := (1, \mathbb{1}(x \geq 0), x/h, x\mathbb{1}(x \geq 0)/h)^\top$ that

$$\begin{aligned} &\sup_{n \in \mathbb{N}} \sup_{x \in [-h,h]} \mathbb{E}\left(\left(V_i^{(m)} Z_i^{(k)} \right)^2 \,\middle|\, X_i = x \right) \\ &= \sup_{n \in \mathbb{N}} \sup_{x \in [-h,h]} u_m(x)^2 \mathbb{E}\left(\left(Z_i^{(k)} \right)^2 \,\middle|\, X_i = x \right) \\ &\leq \sup_{n \in \mathbb{N}} \sup_{x \in [-h,h]} \mathbb{E}\left(\left(Z_i^{(k)} \right)^2 \,\middle|\, X_i = x \right) < \infty. \end{aligned}$$

As result, we obtain

$$\frac{1}{n}\mathbf{V}^\top \mathbf{K}_h \mathbf{Z} = \mathbb{E}(K_h(X_i)V_i Z_i^\top) + O_P\left(\frac{1}{\sqrt{nh}}\right). \tag{6.4}$$

In an analogous way, the third factor results in

$$\frac{1}{n}\mathbf{Z}^\top \mathbf{K}_h \mathbf{V} = \mathbb{E}(K_h(X_i)Z_i V_i^\top) + O_P\left(\frac{1}{\sqrt{nh}}\right). \tag{6.5}$$

The second factor needs some additional work due to having the convergence inside the inverse. First, we start analogously as for the other two factors and state that

$$\left(\frac{1}{n}\mathbf{Z}^\top \mathbf{K}_h \mathbf{Z}\right)^{-1} = \left(\mathbb{E}(K_h(X_i)Z_i Z_i^\top) + O_P\left(\frac{1}{\sqrt{nh}}\right)\right)^{-1}.$$

As we need the asymptotics outside of the inverse, we need to show in addition that

$$\left(\mathbb{E}(K_h(X_i)Z_i Z_i^\top) + O_P\left(\frac{1}{\sqrt{nh}}\right)\right)^{-1} = \mathbb{E}(K_h(X_i)Z_i Z_i^\top)^{-1} + O_P\left(\frac{1}{\sqrt{nh}}\right).$$

This can indeed be done by using Lemma 2.1 (g). The assumptions hold since

$$\mathbb{E}(K_h(X_i)Z_i Z_i^\top)^{-1} = O(1)$$

by assumption (5.6), $\frac{1}{\sqrt{nh}}$ converges to zero and

$$\left(\frac{1}{n}\mathbf{Z}^\top \mathbf{K}_h \mathbf{Z}\right)^{-1} = O_P(1). \tag{6.6}$$

(6.6) is not obvious and, therefore, we want to verify it: By the definition of $O(1)$, there exist $N_1 \in \mathbb{N}$ and $M > 0$ such that

$$\left\|\mathbb{E}(K_h(X_i)Z_i Z_i^\top)^{-1}\right\|_2 \leq \frac{M}{2}$$

for all $n \geq N_1$. Also, we have

$$\frac{1}{n}\mathbf{Z}^\top \mathbf{K}_h \mathbf{Z} \xrightarrow{P} \mathbb{E}(K_h(X_i)Z_i Z_i^\top)$$

leading to

$$\left(\frac{1}{n}\mathbf{Z}^\top \mathbf{K}_h \mathbf{Z}\right)^{-1} \xrightarrow{P} \mathbb{E}(K_h(X_i)Z_i Z_i^\top)^{-1}$$

because the inverse is continuous. Let $\epsilon > 0$. By the definition of convergence in probability, there exists $N_2 \in \mathbb{N}$ such that for all $n \geq N_2$

$$P\left(\left\|\left(\frac{1}{n}\mathbf{Z}^\top \mathbf{K}_h \mathbf{Z}\right)^{-1} - \mathbb{E}(K_h(X_i)Z_i Z_i^\top)^{-1}\right\|_2 > \frac{M}{2}\right) < \epsilon.$$

Taking that together, we obtain for all $n \geq \max(N_1, N_2)$ that

$$
\begin{aligned}
& P\left(\left\|\left(\frac{1}{n}\mathbf{Z}^\top \mathbf{K}_h \mathbf{Z}\right)^{-1}\right\|_2 > M\right) \\
&= P\left(\left\|\left(\frac{1}{n}\mathbf{Z}^\top \mathbf{K}_h \mathbf{Z}\right)^{-1} - \mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1} + \mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\right\|_2 > M\right) \\
&\leq P\left(\left\|\left(\frac{1}{n}\mathbf{Z}^\top \mathbf{K}_h \mathbf{Z}\right)^{-1} - \mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\right\|_2 + \left\|\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\right\|_2 > M\right) \\
&= P\left(\left\|\left(\frac{1}{n}\mathbf{Z}^\top \mathbf{K}_h \mathbf{Z}\right)^{-1} - \mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\right\|_2 > M - \left\|\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\right\|_2\right) \\
&\leq P\left(\left\|\left(\frac{1}{n}\mathbf{Z}^\top \mathbf{K}_h \mathbf{Z}\right)^{-1} - \mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\right\|_2 > \frac{M}{2}\right) \\
&< \epsilon.
\end{aligned}
$$

Hence, (6.6) is verified which leads to

$$
\left(\frac{1}{n}\mathbf{Z}^\top \mathbf{K}_h \mathbf{Z}\right)^{-1} = \mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1} + O_P\left(\frac{1}{\sqrt{nh}}\right). \tag{6.7}
$$

Overall, taking the equations (6.4), (6.5) and (6.7) together, (2) evaluates as following:

$$
\begin{aligned}
& \left(\frac{1}{n}\mathbf{V}^\top \mathbf{K}_h \mathbf{Z}\right)\left(\frac{1}{n}\mathbf{Z}^\top \mathbf{K}_h \mathbf{Z}\right)^{-1}\left(\frac{1}{n}\mathbf{Z}^\top \mathbf{K}_h \mathbf{V}\right) \\
&= \underbrace{\mathbb{E}(K_h(X_i)V_iZ_i^\top)\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\mathbb{E}(K_h(X_i)Z_iV_i^\top)}_{(\triangle)} + o_P(1).
\end{aligned}
$$

Note that we used property (a), (d) and (f) of Lemma 2.1 here together with the fact that all the means are of order $O(1)$ because $\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1} = O(1)$ by assumption and $\mathbb{E}(K_h(X_i)Z_iV_i^\top)^\top = \mathbb{E}(K_h(X_i)V_iZ_i^\top) = O(1)$ by applying Lemma 5.2 as in (5.11). As we now use equation (5.7) taken from Lemma 5.6, which states that expression $(\triangle)$ is of order $O(h^4)$, we obtain

$$
\left(\frac{1}{n}\mathbf{V}^\top \mathbf{K}_h \mathbf{Z}\right)\left(\frac{1}{n}\mathbf{Z}^\top \mathbf{K}_h \mathbf{Z}\right)^{-1}\left(\frac{1}{n}\mathbf{Z}^\top \mathbf{K}_h \mathbf{V}\right) = O(h^4) + o_P(1) = o_P(1)
$$

by applying Lemma 2.1 (a), (c) and (f). Now we work on term (1). Also here, we apply Lemma 6.1 component-wise. It is clear that the necessary assumption (6.1) is satisfied. Afterwards, we apply Lemma 5.4, and at last we use the properties (c) and (f) of Lemma

2.1. This leads to

$$\begin{aligned}\frac{1}{n}\mathbf{V}^\top\mathbf{K}_h\mathbf{V} &= \mathbb{E}(K_h(X_i)V_iV_i^\top) + O_P\left(\frac{1}{\sqrt{nh}}\right)\\ &= f_X(0)\kappa(K) + o_P(1) + O_P\left(\frac{1}{\sqrt{nh}}\right)\\ &= f_X(0)\kappa(K) + o_P(1).\end{aligned}$$

Overall, the calculations for term (1) and (2) result in

$$\frac{1}{n}\mathbf{V}^\top\mathbf{K}_h^{\frac{1}{2}}M\mathbf{K}_h^{\frac{1}{2}}\mathbf{V} = f_X(0)\kappa(K) + o_P(1).$$

Now we have to show the second assertion of the lemma. We start in the same way by splitting up the expression:

$$\begin{aligned}\frac{1}{n}\mathbf{V}^\top\mathbf{K}_h^{\frac{1}{2}}M\mathbf{K}_h^{\frac{1}{2}}\mathbf{r}(h) &= \frac{1}{n}\mathbf{V}^\top\mathbf{K}_h^{\frac{1}{2}}\left(I - \mathbf{K}_h^{\frac{1}{2}}\mathbf{Z}(\mathbf{Z}^\top\mathbf{K}_h\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{K}_h^{\frac{1}{2}}\right)\mathbf{K}_h^{\frac{1}{2}}\mathbf{r}(h)\\ &= \frac{1}{n}\mathbf{V}^\top\mathbf{K}_h\mathbf{r}(h) - \frac{1}{n}\mathbf{V}^\top\mathbf{K}_h\mathbf{Z}(\mathbf{Z}^\top\mathbf{K}_h\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{K}_h\mathbf{r}(h)\\ &= \underbrace{\frac{1}{n}\mathbf{V}^\top\mathbf{K}_h\mathbf{r}(h)}_{(1)} - \underbrace{\left(\frac{1}{n}\mathbf{V}^\top\mathbf{K}_h\mathbf{Z}\right)\left(\frac{1}{n}\mathbf{Z}^\top\mathbf{K}_h\mathbf{Z}\right)^{-1}\left(\frac{1}{n}\mathbf{Z}^\top\mathbf{K}_h\mathbf{r}(h)\right)}_{(2)}.\end{aligned}$$

We do not have to examine expression (1) since it appears in the expression of the assertion. We just have to show that term (2) is of order $o_P\left(\frac{1}{\sqrt{nh}}\right)$. We know that:

- $\frac{1}{n}\mathbf{V}^\top\mathbf{K}_h\mathbf{Z} = \mathbb{E}(K_h(X_i)V_iZ_i^\top) + O_P\left(\frac{1}{\sqrt{nh}}\right) = o(1) + O_P\left(\frac{1}{\sqrt{nh}}\right) = o_P(1)$ whereby the first equation follows from Lemma 6.1 and the second by $\mathbb{E}(K_h(X_i)V_iZ_i^\top) = o(1)$ which for example can be seen in the calculation of (5.11) or by applying Lemma 5.2. The third equation again follows from properties of Lemma 2.1.
- $\left(\frac{1}{n}\mathbf{Z}^\top\mathbf{K}_h\mathbf{Z}\right)^{-1} = \mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1} + O_P\left(\frac{1}{\sqrt{nh}}\right) = O_P(1)$ whereby we used (6.7), assumption (5.6) and again Lemma 2.1.
- $\frac{1}{n}\mathbf{Z}^\top\mathbf{K}_h\mathbf{r}(h) = \mathbb{E}(K_h(X_i)Z_ir_i(h)) + O_P\left(\frac{1}{\sqrt{nh}}\right) = O_P\left(\frac{1}{\sqrt{nh}}\right)$ by applying Lemma 6.1 with A_i set respectively to the components of $Z_ir_i(h)$ and the fact that

  $$\mathbb{E}(K_h(X_i)Z_ir_i(h)) = 0$$

  due to the definition of the residual with $\theta_0(h)$ and $\gamma_0(h)$ being defined as minimizing arguments.

Since property (e) of Lemma 2.1 implies

$$o_P(1)O_P(1)O_P\left(\frac{1}{\sqrt{nh}}\right) = o_P\left(\frac{1}{\sqrt{nh}}\right),$$

we showed that (2) is of order $o_P\left(\frac{1}{\sqrt{nh}}\right)$, leading to the second assertion of the lemma being proved. □

As a result, we can prove the asymptotic normality of the covariate-adjusted estimator now. This is done by the following final lemma of the chapter.

Lemma 6.4. *Let the assumptions in* (6.3) *of Lemma 6.3 hold. Further, let K be a kernel which is compactly supported on $[-1,1]$ with $K^{(4)}, (K^2)^{(0)} < \infty$ and suppose that $\mathbb{E}(K_h(X_i)Z_iZ_i^\top)$ is invertible with $\left\|\mathbb{E}(K_h(X_i)Z_iZ_i^\top)^{-1}\right\|_2 = O(1)$. Also, let f_X be three times continuously differentiable in a neighborhood around zero with $f_X(0) > 0$. Moreover, assume that μ_Z is continuous and three times one-sided differentiable at zero, whereby $\mu_Z(0) = 0$ and $\lim_{x\searrow 0}\mu_Z'(x) = \lim_{x\nearrow 0}\mu_Z'(x) = 0$. Also, suppose that there is $\delta > 0$ such that*

$$\sup_{n\in\mathbb{N}}\sup_{x\in[-h,h]}\mathbb{E}\left(|r_i(h)|^{2+\delta} \mid X_i = x\right) < \infty$$

as well as the existence of finite numbers $\sigma_l, \sigma_r > 0$ such that

$$\begin{aligned}\lim_{n\to\infty}\sup_{\lambda\in[0,1]}|\mathbb{E}\left(r_i(h)^2 \mid X_i = \lambda h\right) - \sigma_r^2| = 0,\\ \lim_{n\to\infty}\sup_{\lambda\in[0,1]}|\mathbb{E}\left(r_i(h)^2 \mid X_i = -\lambda h\right) - \sigma_l^2| = 0.\end{aligned} \tag{6.8}$$

Set $w = \left(\left[(f_X(0)\kappa(K))^{-1}\right]_{2\cdot}\right)^\top$ and define

$$\mathcal{S}_n^2 = \frac{1}{h}\mathbb{E}\left(K\left(\frac{X_i}{h}\right)^2(w^\top V_i)^2 r_i(h)^2\right).$$

Then,

$$\sqrt{\frac{nh}{\mathcal{S}_n^2}}\left(\hat{\tau}_h - \theta_0^{(2)}\right) \xrightarrow{d} \mathcal{N}(0,1).$$

Proof. Denote by

$$M = I - \mathbf{K}_h^{\frac{1}{2}}\mathbf{Z}(\mathbf{Z}^\top\mathbf{K}_h\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{K}_h^{\frac{1}{2}}$$

the projection matrix. First note that

$$M\mathbf{K}_h^{\frac{1}{2}}\mathbf{Z} = 0 \tag{6.9}$$

by using the definition of M. Now, define $\bar{X} := \mathbf{K}_h^{\frac{1}{2}}\begin{pmatrix}\mathbf{V} & \mathbf{Z}\end{pmatrix}$ and $\bar{Y} := \mathbf{K}_h^{\frac{1}{2}}\mathbf{Y}$. The linear least squares problem

$$(\hat{\theta}_n, \hat{\gamma}_n) = \underset{(\theta,\gamma)\in\mathbb{R}^{p+4}}{\operatorname{argmin}}\sum_{i=1}^n K_h(X_i)\left(Y_i - V_i^\top\theta - Z_i^\top\gamma\right)^2$$

can be written as

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^{p+4}}{\operatorname{argmin}}\left\|\bar{Y} - \bar{X}\beta\right\|_2^2,$$

which has the solution

$$\hat{\beta} = (\hat{\theta}_n, \hat{\gamma}_n) = (\bar{X}^\top \bar{X})^{-1} \bar{X}^\top \bar{Y}.$$

We calculate this expression for $\hat{\beta}$. In fact, as we have

$$\bar{X}^\top \bar{X} = \begin{pmatrix} \mathbf{V}^\top \mathbf{K}_h \mathbf{V} & \mathbf{V}^\top \mathbf{K}_h \mathbf{Z} \\ \mathbf{Z}^\top \mathbf{K}_h \mathbf{V} & \mathbf{Z}^\top \mathbf{K}_h \mathbf{Z} \end{pmatrix},$$

we obtain by applying the well-known formula for the inverse of block matrices that

$$(\bar{X}^\top \bar{X})^{-1} = \begin{pmatrix} R^{-1} & -R^{-1} \mathbf{V}^\top \mathbf{K}_h \mathbf{Z} (\mathbf{Z}^\top \mathbf{K}_h \mathbf{Z})^{-1} \\ * & * \end{pmatrix} \tag{6.10}$$

whereby

$$\begin{aligned} R &:= \mathbf{V}^\top \mathbf{K}_h \mathbf{V} - \mathbf{V}^\top \mathbf{K}_h \mathbf{Z} (\mathbf{Z}^\top \mathbf{K}_h \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{K}_h \mathbf{V} \\ &= \mathbf{V}^\top \mathbf{K}_h^{\frac{1}{2}} (I - \mathbf{K}_h^{\frac{1}{2}} \mathbf{Z} (\mathbf{Z}^\top \mathbf{K}_h \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{K}_h^{\frac{1}{2}}) \mathbf{K}_h^{\frac{1}{2}} \mathbf{V} \\ &= \mathbf{V}^\top \mathbf{K}_h^{\frac{1}{2}} M \mathbf{K}_h^{\frac{1}{2}} \mathbf{V}. \end{aligned}$$

Furthermore,

$$\bar{X}^\top \bar{Y} = \begin{pmatrix} \mathbf{V}^\top \mathbf{K}_h \mathbf{Y} \\ \mathbf{Z}^\top \mathbf{K}_h \mathbf{Y} \end{pmatrix}. \tag{6.11}$$

Taking (6.10) and (6.11) together delivers

$$\begin{aligned} \hat{\theta}_n &= (\mathbf{V}^\top \mathbf{K}_h^{\frac{1}{2}} M \mathbf{K}_h^{\frac{1}{2}} \mathbf{V})^{-1} (\mathbf{V}^\top K_h \mathbf{Y} - \mathbf{V}^\top \mathbf{K}_h \mathbf{Z} (\mathbf{Z}^\top \mathbf{K}_h \mathbf{Z})^{-1} \mathbf{Z}^\top K_h \mathbf{Y}) \\ &= (\mathbf{V}^\top \mathbf{K}_h^{\frac{1}{2}} M \mathbf{K}_h^{\frac{1}{2}} \mathbf{V})^{-1} (\mathbf{V}^\top \mathbf{K}_h^{\frac{1}{2}} (I - \mathbf{K}_h^{\frac{1}{2}} \mathbf{Z} (\mathbf{Z}^\top \mathbf{K}_h \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{K}_h^{\frac{1}{2}}) \mathbf{K}_h^{\frac{1}{2}} \mathbf{Y}) \\ &= (\mathbf{V}^\top \mathbf{K}_h^{\frac{1}{2}} M \mathbf{K}_h^{\frac{1}{2}} \mathbf{V})^{-1} \mathbf{V}^\top \mathbf{K}_h^{\frac{1}{2}} M \mathbf{K}_h^{\frac{1}{2}} \mathbf{Y}. \end{aligned}$$

Considering that

$$\mathbf{Y} = \mathbf{r}(h) + \mathbf{V}\theta_0(h) + \mathbf{Z}\gamma_0(h)$$

finally delivers the representation

$$\begin{aligned} \hat{\theta}_n =& \ (\mathbf{V}^\top \mathbf{K}_h^{\frac{1}{2}} M \mathbf{K}_h^{\frac{1}{2}} \mathbf{V})^{-1} \mathbf{V}^\top \mathbf{K}_h^{\frac{1}{2}} M \mathbf{K}_h^{\frac{1}{2}} (\mathbf{r}(h) + \mathbf{V}\theta_0(h) + \mathbf{Z}\gamma_0(h)) \\ =& \ (\mathbf{V}^\top \mathbf{K}_h^{\frac{1}{2}} M \mathbf{K}_h^{\frac{1}{2}} \mathbf{V})^{-1} \mathbf{V}^\top \mathbf{K}_h^{\frac{1}{2}} M \mathbf{K}_h^{\frac{1}{2}} (\mathbf{Z}\gamma_0(h) + \mathbf{r}(h)) \\ &+ \underbrace{(\mathbf{V}^\top \mathbf{K}_h^{\frac{1}{2}} M \mathbf{K}_h^{\frac{1}{2}} \mathbf{V})^{-1} \mathbf{V}^\top \mathbf{K}_h^{\frac{1}{2}} M \mathbf{K}_h^{\frac{1}{2}} \mathbf{V}}_{=I} \theta_0(h) \\ =& \ \theta_0(h) + (\mathbf{V}^\top \mathbf{K}_h^{\frac{1}{2}} M \mathbf{K}_h^{\frac{1}{2}} \mathbf{V})^{-1} \mathbf{V}^\top \mathbf{K}_h^{\frac{1}{2}} M \mathbf{K}_h^{\frac{1}{2}} (\mathbf{Z}\gamma_0(h) + \mathbf{r}(h)) \\ \overset{(6.9)}{=}& \ \theta_0(h) + (\mathbf{V}^\top \mathbf{K}_h^{\frac{1}{2}} M \mathbf{K}_h^{\frac{1}{2}} \mathbf{V})^{-1} \mathbf{V}^\top \mathbf{K}_h^{\frac{1}{2}} M \mathbf{K}_h^{\frac{1}{2}} \mathbf{r}(h). \end{aligned} \tag{6.12}$$

By applying Lemma 6.3, we obtain

$$\mathbf{V}^\top \mathbf{K}_h^{\frac{1}{2}} M \mathbf{K}_h^{\frac{1}{2}} \mathbf{r}(h) = \frac{1}{n} \mathbf{V}^\top \mathbf{K}_h \mathbf{r}(h) + o_P\left(\frac{1}{\sqrt{nh}}\right) \tag{6.13}$$

and

$$(\mathbf{V}^\top \mathbf{K}_h^{\frac{1}{2}} M \mathbf{K}_h^{\frac{1}{2}} \mathbf{V})^{-1} = (f_X(0)\kappa(K) + o_P(1))^{-1} \overset{(*)}{=} (f_X(0)\kappa(K))^{-1} + o_P(1). \tag{6.14}$$

Note that $(*)$ follows from Lemma 2.1 (h) since all the necessary assumptions are true: $f_X(0)\kappa(K)$ is invertible, $(f_X(0)\kappa(K))^{-1} = O(1)$ and

$$(f_X(0)\kappa(K) + o_P(1))^{-1} = O_P(1), \tag{6.15}$$

whereby (6.15) follows from $(f_X(0)\kappa(K) + o_P(1))^{-1} \xrightarrow{P} (f_X(0)\kappa(K))^{-1} = O(1)$ which allows us to conclude with the same reasoning used to show (6.6). Hence, inserting (6.13) and (6.14) into (6.12), gives

$$\hat{\theta}_n = \theta_0 + ((f_X(0)\kappa(K))^{-1} + o_P(1)) \left(\frac{1}{n} \mathbf{V}^\top \mathbf{K}_h \mathbf{r}(h) + o_P\left(\frac{1}{\sqrt{nh}}\right)\right). \tag{6.16}$$

We know by Lemma 6.2, whose assumption (6.2) is satisfied by (6.8), together with Lemma 2.1 (e), that

$$o_P(1)\frac{1}{n} \mathbf{V}^\top \mathbf{K}_h \mathbf{r}(h) = o_P(1) O_P\left(\frac{1}{\sqrt{nh}}\right) = o_P\left(\frac{1}{\sqrt{nh}}\right).$$

Therefore, by also using Lemma 2.1 (d), (6.16) expands to

$$\hat{\theta}_n = \theta_0 + (f_X(0)\kappa(K))^{-1} \frac{1}{n} \mathbf{V}^\top \mathbf{K}_h \mathbf{r}(h) + (f_X(0)\kappa(K))^{-1} o_P\left(\frac{1}{\sqrt{nh}}\right) + o_P\left(\frac{1}{\sqrt{nh}}\right),$$

which, overall, leads to

$$\sqrt{nh}\left(\hat{\theta}_n - \theta_0\right) = (f_X(0)\kappa(K))^{-1} \sqrt{nh} \frac{1}{n} \mathbf{V}^\top \mathbf{K}_h \mathbf{r}(h) + o_P(1). \tag{6.17}$$

Since we are interested in the estimator of the average treatment effect, we need to study the second row of that expression. Please recall the definition of w given in the assertion of the lemma. Then, the second row of (6.17) can be written as

$$\sqrt{nh}\left(\hat{\tau}_h - \theta_0^{(2)}\right) = \sqrt{nh} \frac{1}{n} w^\top \mathbf{V}^\top \mathbf{K}_h \mathbf{r}(h) + o_P(1) = \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} K\left(\frac{X_i}{h}\right) w^\top V_i r_i(h) + o_P(1).$$

Next, we want to use Lyapunov's Central Limit Theorem, which is stated in Proposition 2, to argue that the result follows. Define $\nu(X_i/h) := w^\top V_i$. Then, the independent sequence

of random variables is considered to be

$$\left(\frac{1}{\sqrt{nh}}K\left(\frac{X_i}{h}\right)\nu\left(\frac{X_i}{h}\right)r_i(h)\right)_{i=1,\dots,n}$$

whereby each of this random variables has expectation zero due to the definition of the residual. We calculate the variance $\mathcal{S}_n^2$ of the sum of the variables now:

$$\begin{aligned}\mathcal{S}_n^2 &= \operatorname{Var}\left(\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}K\left(\frac{X_i}{h}\right)\nu\left(\frac{X_i}{h}\right)r_i(h)\right)\\ &= \frac{1}{nh}\sum_{i=1}^{n}\mathbb{E}\left(K\left(\frac{X_i}{h}\right)^2\nu\left(\frac{X_i}{h}\right)^2 r_i(h)^2\right)\\ &= \frac{1}{h}\mathbb{E}\left(K\left(\frac{X_i}{h}\right)^2\nu\left(\frac{X_i}{h}\right)^2 r_i(h)^2\right).\end{aligned}$$

Now we need to state that the required condition

$$\lim_{n\to\infty}\frac{1}{\mathcal{S}_n^{2+\delta}}\sum_{i=1}^{n}\frac{1}{\left(\sqrt{nh}\right)^{2+\delta}}\mathbb{E}\left(K\left(\frac{X_i}{h}\right)^{2+\delta}\left|\nu\left(\frac{X_i}{h}\right)\right|^{2+\delta}|r_i(h)|^{2+\delta}\right)=0 \tag{6.18}$$

for Lyapunov's CLT is true. Indeed, there are constants $C_1, C_2 > 0$ and $\delta > 0$ such that

$$\frac{1}{h}\mathbb{E}\left(K\left(\frac{X_i}{h}\right)^2\nu\left(\frac{X_i}{h}\right)^2 r_i(h)^2\right)\xrightarrow[n\to\infty]{} C_1, \tag{6.19}$$

$$\frac{1}{h}\mathbb{E}\left(K\left(\frac{X_i}{h}\right)^{2+\delta}\left|\nu\left(\frac{X_i}{h}\right)\right|^{2+\delta}|r_i(h)|^{2+\delta}\right)\le C_2. \tag{6.20}$$

The convergence in (6.19) can be seen by

$$\begin{aligned}&\lim_{n\to\infty}\frac{1}{h}\mathbb{E}\left(K\left(\frac{X_i}{h}\right)^2\nu\left(\frac{X_i}{h}\right)^2 r_i(h)^2\right)\\ =\ &\lim_{n\to\infty}\frac{1}{h}\mathbb{E}\left(K\left(\frac{X_i}{h}\right)^2\nu\left(\frac{X_i}{h}\right)^2\mathbb{E}(r_i(h)^2\mid X_i)\right)\\ =\ &\lim_{n\to\infty}\int_{-\infty}^{\infty}K(y)^2\nu(y)^2\mathbb{E}(r_i(h)^2\mid X_i=yh)f_X(yh)\,dy\\ =\ &\int_{-1}^{1}K(y)^2\nu(y)^2\lim_{n\to\infty}\mathbb{E}(r_i(h)^2\mid X_i=yh)f_X(yh)\,dy\\ =\ &\int_{-1}^{0}K(y)^2\nu(y)^2\lim_{n\to\infty}\mathbb{E}(r_i(h)^2\mid X_i=yh)f_X(yh)\,dy\\ &+\int_{0}^{1}K(y)^2\nu(y)^2\lim_{n\to\infty}\mathbb{E}(r_i(h)^2\mid X_i=yh)f_X(yh)\,dy\end{aligned}$$

whereby we used the Theorem of Dominated Convergence to move the limit inside the

integral. For both $1 \geq y > 0$ and $-1 \leq y < 0$, $\lim_{n\to\infty} \mathbb{E}(r_i(h)^2 \mid X_i = yh)$ exists directly due to the assumptions stated in (6.8). Thus, the above expression converges to some $C_1 > 0$.
The upper bound in (6.20) is a direct consequence of the assumption

$$\sup_{n\in\mathbb{N}} \sup_{x\in[-h,h]} \mathbb{E}(|r_i(h)|^{2+\delta} \mid X_i = x) < \infty$$

since

$$\begin{aligned}
&\frac{1}{h}\mathbb{E}\left(K\left(\frac{X_i}{h}\right)^{2+\delta}\left|\nu\left(\frac{X_i}{h}\right)\right|^{2+\delta}|r_i(h)|^{2+\delta}\right) \\
&= \int_{-1}^{1} K(y)^{2+\delta}|\nu(y)|^{2+\delta}\mathbb{E}(|r_i(h)|^{2+\delta} \mid X_i = yh) f_X(yh)\,dy.
\end{aligned}$$

Combining (6.19) and (6.20) gives

$$\begin{aligned}
&\frac{1}{\mathcal{S}_n^{2+\delta}} \sum_{i=1}^{n} \frac{1}{\left(\sqrt{nh}\right)^{2+\delta}} \mathbb{E}\left(K\left(\frac{X_i}{h}\right)^{2+\delta}\left|\nu\left(\frac{X_i}{h}\right)\right|^{2+\delta}|r_i(h)|^{2+\delta}\right) \\
&= \sum_{i=1}^{n} \frac{n^{-\frac{2+\delta}{2}} h^{-\frac{2+\delta}{2}} \mathbb{E}\left(K\left(\frac{X_i}{h}\right)^{2+\delta}\left|\nu\left(\frac{X_i}{h}\right)\right|^{2+\delta}|r_i(h)|^{2+\delta}\right)}{\mathcal{S}_n^{2+\delta}} \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{n^{-\frac{\delta}{2}} h^{-\frac{\delta}{2}} \frac{1}{h}\mathbb{E}\left(K\left(\frac{X_i}{h}\right)^{2+\delta}\left|\nu\left(\frac{X_i}{h}\right)\right|^{2+\delta}|r_i(h)|^{2+\delta}\right)}{\mathcal{S}_n^{2+\delta}} \\
&\leq \frac{C_2(nh)^{-\frac{\delta}{2}}}{\left(\frac{C_1}{2}\right)^{\frac{2+\delta}{2}}} \xrightarrow[n\to\infty]{} 0 \quad (\text{since } nh \to \infty)
\end{aligned}$$

which is condition (6.18) we wanted to show. Hence, as Lyapunov's CLT stated in Proposition 2 can be applied, we obtain

$$\frac{1}{\sqrt{nh\mathcal{S}_n^2}} \sum_{i=1}^{n} K\left(\frac{X_i}{h}\right)\nu\left(\frac{X_i}{h}\right) r_i(h) \xrightarrow{d} \mathcal{N}(0,1).$$

This shows the assertion. □

7 Asymptotic Normality of the Estimator

This chapter aims to prove the main theorem of this thesis, stating that the estimator of the average treatment effect is asymptotically normally distributed. As several conditions are required so that this holds, the next section will briefly cover all necessary assumptions. After that, the final theorem will be stated and proved, followed by two lemmas that cover the rather technical calculations which show the representations of the bias and variance used within the theorem's assertion. Finally, we will compare these results to the bias and variance of the baseline estimator without covariates.

7.1 Assumptions

In order to state the asymptotic normality of the covariate-adjusted estimator of the average treatment effect, we need to formulate several assumptions that all need to be satisfied.

Assumption 1 (Convergence of bandwidth)**.** *Let the bandwidth be a sequence* $(h_n)_{n\in\mathbb{N}}$ *with* $h_n > 0$ *for all* $n \in \mathbb{N}$ *such that* $h \to 0$ *and* $nh \to \infty$ *for* $n \to \infty$*.*

Assumption 2 (Kernel function)**.** *Let* $K : \mathbb{R} \to \mathbb{R}$ *be a non-negative, integrable and symmetric function with* $K^{(4)}, \left(K^2\right)^{(2)} < \infty$ *which integrates to one and is supported on* $[-1, 1]$*.*

Assumption 3 (Differentiability)**.** *Let* μ_Z *be continuous. Further, let the conditional expectations* μ_Z *and* μ_Y *be three times one-sided differentiable at zero, that is being three times differentiable on* $(-\infty, 0)$ *and* $(0, \infty)$ *such that the left- and right-sided limit in zero of* μ_Z *and* μ_Y *and its derivatives up to order three exist. Also, let* $\mu_{ZZ^\top}$ *and* μ_{ZY} *be one time one-sided differentiable in zero.*

In order to state the next assumptions, the following two definitions regarding an adjusted version of the covariates and a covariate-adjusted outcome are necessary.

Definition 7.1. *Define the matrix* $M_n \in \mathbb{R}^{4\times p}$ *as*

$$M_n := \begin{pmatrix} \mu_Z(0) & \mathbf{0} & h\mu'_{Z-} & h\left(\mu'_{Z+} - \mu'_{Z-}\right) \end{pmatrix}^\top.$$

Definition 7.2 (Adjusted variables)**.** *Define*

$$\tilde{Z}_i := Z_i - M_n^\top V_i \quad \text{and} \quad \tilde{Y}_i := Y_i - Z_i^\top \gamma_n$$

with $\gamma_n := \left(\sigma^2_{Z-} + \sigma^2_{Z+}\right)^{-1} \left(\sigma^2_{ZY-} + \sigma^2_{ZY+}\right)$*.*

Please note that replacing Z_i by $\tilde{Z}_i$ in the definition of $\tilde{Y}_i$ and γ_n would deliver the same representation of the bias and variance as provided in (7.1). This can be verified similarly to statements contained in the calculation of that representations, particularly in (7.7) and (7.8) in combination with (7.4). Now we can continue with the next assumption.

Assumption 4 (Invertibility). *Let* $\mathbb{E}\left(K_h(X_i)\tilde{Z}_i\tilde{Z}_i^\top\right)$ *be invertible and*

$$\left\|\mathbb{E}\left(K_h(X_i)\tilde{Z}_i\tilde{Z}_i^\top\right)^{-1}\right\|_2 = O(1).$$

Assumption 5 (Density). *Let the probability density function* f_X *of* X_i *be three times continuously differentiable in a neighborhood around zero with* $f_X(0) > 0$.

Assumption 6 (Uniform boundedness). *Let the following statements be true for all* $k, l \in \{1, ..., p\}$

$$\sup_{n\in\mathbb{N}}\sup_{x\in[-h,h]}\mathbb{E}\left(\left(\tilde{Z}_i^{(k)}r_i(h)\right)^2 \;\middle|\; X_i = x\right) < \infty,$$
$$\sup_{n\in\mathbb{N}}\sup_{x\in[-h,h]}\mathbb{E}\left(\left(\tilde{Z}_i^{(k)}\tilde{Z}_i^{(l)}\right)^2 \;\middle|\; X_i = x\right) < \infty$$

and let $\delta > 0$ *such that*

$$\sup_{n\in\mathbb{N}}\sup_{x\in[-h,h]}\mathbb{E}\left(|r_i(h)|^{2+\delta} \mid X_i = x\right) < \infty.$$

Assumption 7 (Equi-continuity). *Let* $\sigma_l, \sigma_r > 0$ *be finite numbers such that*

$$\lim_{n\to\infty}\sup_{\lambda\in[0,1]}\left|\mathbb{E}\left(r_i(h)^2 \mid X_i = \lambda h\right) - \sigma_r^2\right| = 0,$$
$$\lim_{n\to\infty}\sup_{\lambda\in[0,1]}\left|\mathbb{E}\left(r_i(h)^2 \mid X_i = -\lambda h\right) - \sigma_l^2\right| = 0.$$

7.2 Main Theorem

The representation of the bias and the variance that are used in the theorem include two coefficients being dependent of the kernel function. After defining those, we will finally state the theorem.

Definition 7.3. *Define*

$$\mathcal{C}_\mathcal{B} := \frac{2\left(K_+^{(2)}\right)^2 - 2K_+^{(1)}K_+^{(3)}}{K_+^{(2)} - 2\left(K_+^{(1)}\right)^2}$$

and

$$\mathcal{C}_\mathcal{S} := \frac{\left(K^2\right)^{(0)}\left(K_+^{(2)}\right)^2 + \left(K^2\right)_+^{(2)}\left(K_+^{(1)}\right)^2 - 2\left(K^2\right)_+^{(1)}K_+^{(2)}K_+^{(1)}}{\left[\left(K_+^{(1)}\right)^2 - \frac{1}{2}K_+^{(2)}\right]^2}.$$

Theorem 1. *Suppose that the assumptions of Section 7.1 hold. There are sequences $\mathcal{B}_n$ and $\mathcal{S}_n$ with*

$$\begin{aligned} \mathcal{B}_n &= \frac{\mathcal{C}_{\mathcal{B}}}{2}\left(\mu''_{\tilde{Y}+} - \mu''_{\tilde{Y}-}\right) + o(1) \\ \mathcal{S}_n^2 &= \frac{\mathcal{C}_{\mathcal{S}}}{f_X(0)}\left(\sigma^2_{\tilde{Y}+} + \sigma^2_{\tilde{Y}-}\right) + o(1) \end{aligned} \tag{7.1}$$

such that the following statement is true:

$$\frac{\sqrt{nh}(\hat{\tau}_h - \tau_Y - h^2\mathcal{B}_n)}{\mathcal{S}_n} \xrightarrow{d} \mathcal{N}(0,1).$$

The proof of the theorem consists of two main steps. The first one is to show a representation of the bias. Indeed, this was the result of Chapter 5. In the second main step, we show the asymptotic normality in the covariate-adjusted case. Indeed, this is exactly what was done in Chapter 6. In fact, combining Chapter 5 and 6 gives us the proof of the main theorem of this thesis. All we have to do is to verify the necessary conditions and requirements. Besides, just some extra effort is required in order to obtain the above representation of the bias and variance as stated in the assertion.

Proof of Theorem 1. We start by showing that it is enough to consider the case when μ_Z is continuously differentiable. Note that the differentiablity and continuity statements about μ_Z in Assumption 3 are necessary throughout the next few calculations on $\mu_{\tilde{Z}}$. First, we have

$$\begin{aligned} \mu_{\tilde{Z}}(x) &= \mathbb{E}(\tilde{Z}_i \mid X_i = x) \\ &= \mathbb{E}\left(Z_i \mid X_i = x\right) - \mathbb{E}\left(M_n^\top V_i \mid X_i = x\right) \\ &= \mathbb{E}(Z_i \mid X_i = x) - \mathbb{E}\left(\mu_Z(0) + X_i\mu'_{Z-} + T_iX_i\left(\mu'_{Z+} - \mu'_{Z-}\right) \mid X_i = x\right) \\ &= \mu_Z(x) - \mu_Z(0) - x\mu'_{Z-} - x\mathbb{1}(x \geq 0)\left(\mu'_{Z+} - \mu'_{Z-}\right). \end{aligned}$$

Now we can state the following: For $x > 0$ we have

$$\begin{aligned} \mu_{\tilde{Z}}(x) &= \mu_Z(x) - \mu_Z(0) - x\mu'_{Z-} - x\left(\mu'_{Z+} - \mu'_{Z-}\right) \\ &= \mu_Z(x) - \mu_Z(0) - x\mu'_{Z+}, \\ \mu'_{\tilde{Z}}(x) &= \mu'_Z(x) - \mu'_{Z+}, \end{aligned}$$

which implies

$$\lim_{x \searrow 0} \mu'_{\tilde{Z}}(x) = \mu'_{Z+} - \mu'_{Z+} = 0. \tag{7.2}$$

On the other hand, we obtain for $x < 0$

$$\mu_{\tilde{Z}}(x) = \mu_Z(x) - \mu_Z(0) - x\mu'_{Z-},$$
$$\mu'_{\tilde{Z}}(x) = \mu'_Z(x) - \mu'_{Z-},$$

hence

$$\lim_{x \nearrow 0} \mu'_{\tilde{Z}}(x) = \mu'_{Z-} - \mu'_{Z-} = 0. \tag{7.3}$$

Therefore, we can continuously extend $\mu'_{\tilde{Z}}$ at zero by setting $\mu'_{\tilde{Z}}(0) = 0$. Also, we have

$$\mu_{\tilde{Z}}(0) = \mu_Z(0) - \mu_Z(0) - 0\mu'_{Z-} - 0\mathbb{1}(x \geq 0)\left(\mu'_{Z+} - \mu'_{Z-}\right) = 0. \tag{7.4}$$

In particular, Assumption 3 implies that also $\mu_{\tilde{Z}}$ is continuous and three times one-sided differentiable at zero. Similarly, we can show that $\mu_{\tilde{Z}\tilde{Z}^\top}$ and $\mu_{\tilde{Z}Y}$ are one time one-sided differentiable at zero.

By defining an estimator which is based on V_i and $\tilde{Z}_i$, namely

$$\left(\check{\theta}_n(h), \check{\gamma}_n(h)\right) = \underset{(\theta,\gamma)\in\mathbb{R}^{p+4}}{\operatorname{argmin}} \sum_{i=1}^{n} K_h(X_i)\left(Y_i - V_i^\top\theta - \tilde{Z}_i{}^\top\gamma\right)^2,$$

we can compare it with our normal estimator based on V_i and Z_i:

$$\begin{aligned}\left(\hat{\theta}_n(h), \hat{\gamma}_n(h)\right) &= \underset{(\theta,\gamma)\in\mathbb{R}^{p+4}}{\operatorname{argmin}} \sum_{i=1}^{n} K_h(X_i)\left(Y_i - V_i^\top\theta - Z_i^\top\gamma\right)^2 \\ &= \underset{(\theta,\gamma)\in\mathbb{R}^{p+4}}{\operatorname{argmin}} \sum_{i=1}^{n} K_h(X_i)\left(Y_i - V_i^\top(\theta + M_n\gamma) - \tilde{Z}_i{}^\top\gamma\right)^2.\end{aligned}$$

As it is obvious that the minimizing parameters have to be identical, we conclude in the case of uniqueness of these parameters

$$\check{\gamma}_n(h) = \hat{\gamma}_n(h) \quad \text{and} \quad \check{\theta}_n(h) = \hat{\theta}_n(h) + M_n\hat{\gamma}_n(h).$$

In the case when multiple minimizing parameters exist, we just state that there are parameters such that the above identification holds. In an analogous way, we also have

$$\left(\check{\theta}_0(h), \check{\gamma}_0(h)\right) := \underset{(\theta,\gamma)\in\mathbb{R}^{p+4}}{\operatorname{argmin}} \mathbb{E}\left(K_h(X_i)\left(Y_i - V_i^\top\theta - \tilde{Z}_i{}^\top\gamma\right)^2\right)$$

and

$$(\theta_0(h), \gamma_0(h)) = \underset{(\theta,\gamma)\in\mathbb{R}^{p+4}}{\operatorname{argmin}} \mathbb{E}\left(K_h(X_i)\left(Y_i - V_i^\top(\theta + M_n\gamma) - \tilde{Z}_i{}^\top\gamma\right)^2\right).$$

By providing the same argument as above, we obtain

$$\check{\gamma}_0(h) = \gamma_0(h) \quad \text{and} \quad \check{\theta}_0(h) = \theta_0(h) + M_n\gamma_0(h).$$

As we are only interested in the effect of the treatment variable, we proceed by calculating the second entry of those parameters:

$$\check{\theta}_n^{(2)}(h) = \hat{\theta}_n^{(2)}(h) + (M_n)_2 \hat{\gamma}_n(h) = \hat{\theta}_n^{(2)}(h),$$

whereby the second equality follows from the fact that the second row of M_n is zero by definition. We can now substitute the estimators in the expression which we want to examine asymptotically:

$$\sqrt{\frac{nh}{\mathcal{S}_n^2}} \left(\hat{\theta}_n^{(2)}(h) - \check{\theta}_0^{(2)}(h) \right) = \sqrt{\frac{nh}{\mathcal{S}_n^2}} \left(\check{\theta}_n^{(2)}(h) - \check{\theta}_0^{(2)}(h) \right). \tag{7.5}$$

We want to apply Lemma 5.7 with $Z_i = \tilde{Z}_i$ now. Indeed, all necessary conditions are satisfied due to Assumption 1 to 5 and the considerations in (7.2), (7.3) and (7.4). Indeed, Lemma 5.7 gives us $\check{\theta}_0^{(2)} = \tau_Y + h^2 \check{\mathcal{B}}_n$ whereby

$$\check{\mathcal{B}}_n = \frac{\mathcal{C}_{\mathcal{B}}}{2} \left(\mu''_{Y+} - \mu''_{Y-} - \sum_{k=1}^{p} \left(\mu''_{\tilde{Z}^{(k)}+} - \mu''_{\tilde{Z}^{(k)}-} \right) \check{\beta}_n^{(k)} \right) + o(1)$$

with

$$\check{\beta}_n = \mathbb{E}\left(K_h(X_i) \tilde{Z}_i \tilde{Z}_i^\top \right)^{-1} \mathbb{E}\left(K_h(X_i) \tilde{Z}_i Y_i \right).$$

By Lemma 7.1, which can be found below and calculates the representation for the bias, we can conclude $\check{\mathcal{B}}_n = \mathcal{B}_n$ leading to

$$\check{\theta}_0^{(2)} = \tau_Y + h^2 \mathcal{B}_n.$$

Also, we know that $\hat{\theta}_n^{(2)} = \hat{\tau}_h$ by definition. Inserting these expressions in (7.5) shows that it is enough to examine the asymptotic behavior of

$$\sqrt{\frac{nh}{\mathcal{S}_n^2}} \left(\check{\theta}_n^{(2)}(h) - \check{\theta}_0^{(2)}(h) \right)$$

in order to prove the assertion of the theorem.

We have

$$\begin{aligned} \check{r}_i(h) &= Y_i - V_i^\top \check{\theta}_0(h) - \tilde{Z}_i^\top \check{\gamma}_0(h) \\ &= Y_i - V_i^\top \left(\theta_0(h) + M_n \gamma_0(h) \right) - \left(Z_i - M_n^\top V_i \right)^\top \gamma_0(h) \\ &= Y_i - V_i^\top \theta_0(h) - Z_i \gamma_0(h) \\ &= r_i(h). \end{aligned} \tag{7.6}$$

As a consequence, we can replace $r_i(h)$ by $\check{r}_i(h)$ in Assumption 6 and 7 and they still hold. Thus, Lemma 6.4 can be applied with $Z_i = \tilde{Z}_i$ since its assumptions are satisfied by using

Assumption 1 to 7 from Section 7.1. This gives us that

$$\sqrt{\frac{nh}{\check{\mathcal{S}}_n^2}}\left(\check{\theta}_n^{(2)}(h) - \check{\theta}_0^{(2)}(h)\right) \xrightarrow{d} \mathcal{N}(0,1)$$

whereby

$$\check{\mathcal{S}}_n^2 = \frac{1}{h}\mathbb{E}\left(K\left(\frac{X_i}{h}\right)^2 (w^\top V_i)^2 \check{r}_i(h)^2\right).$$

From Lemma 7.2, which can be found below and calculates the representation of the variance, we know that $\check{\mathcal{S}}_n^2 = \mathcal{S}_n^2$, directly leading to

$$\frac{\sqrt{nh}(\hat{\tau}_h - \tau_Y - h^2\mathcal{B}_n)}{\mathcal{S}_n} \xrightarrow{d} \mathcal{N}(0,1)$$

by the above considerations. This proves the assertion of the theorem. □

The subsequent two lemmas cover the calculation of the representation of the bias and the variance used in the above theorem. Mostly, it relies on rather technical computations and already shown representations from the previous chapters.

Lemma 7.1 (Computation for the bias)**.** *Let the assumptions of Section 7.1 hold. Then*

$$\mu''_{Y+} - \mu''_{Y-} - \left(\mu''_{\tilde{Z}+} - \mu''_{\tilde{Z}-}\right)^\top \check{\beta}_n = \mu''_{\tilde{Y}+} - \mu''_{\tilde{Y}-} + o(1)$$

whereby

$$\check{\beta}_n = \mathbb{E}\left(K_h(X_i)\tilde{Z}_i\tilde{Z}_i^\top\right)^{-1}\mathbb{E}\left(K_h(X_i)\tilde{Z}_i Y_i\right).$$

Proof. First, we know that

$$\mu''_Z = \mu''_{\tilde{Z}}. \tag{7.7}$$

For this reason, we can also proof

$$\mu''_{Y+} - \mu''_{Y-} - \left(\mu''_{Z+} - \mu''_{Z-}\right)^\top \check{\beta}_n = \mu''_{\tilde{Y}+} - \mu''_{\tilde{Y}-} + o(1)$$

in order to show the lemma. Since we know by definition that for $x \neq 0$

$$\begin{aligned}
\mu''_{\tilde{Y}}(x) &= \mu''_{Y - Z^\top\gamma_n}(x) \\
&= \frac{\mathrm{d}^2}{\mathrm{d}x^2}\,\mathbb{E}\left(Y_i - Z_i^\top\gamma_n \mid X_i = x\right) \\
&= \frac{\mathrm{d}^2}{\mathrm{d}x^2}\,\mathbb{E}(Y_i \mid X_i = x) - \frac{\mathrm{d}^2}{\mathrm{d}x^2}\,\mathbb{E}(Z_i^\top \mid X_i = x)\gamma_n \\
&= \mu''_Y - \mu''_{Z^\top}\gamma_n,
\end{aligned}$$

we obtain

$$\mu''_{\tilde{Y}+} - \mu''_{\tilde{Y}-} = \mu''_{Y+} - \mu''_{Y-} - \left(\mu''_{Z+} - \mu''_{Z-}\right)^\top \gamma_n.$$

Therefore it is enough to show that

$$\left(\mu''_{Z+} - \mu''_{Z-}\right)^\top \left(\check{\beta}_n - \gamma_n\right) = o(1).$$

Since the first factor is constant, we just have to evaluate the convergence of the second factor. Indeed,

$$\begin{aligned}\left\|\check{\beta}_n - \gamma_n\right\|_2 &= \left\|\mathbb{E}\left(K_h(X_i)\tilde{Z}_i\tilde{Z}_i^\top\right)^{-1}\mathbb{E}\left(K_h(X_i)\tilde{Z}_i\left(Y_i - \tilde{Z}_i^\top\gamma_n\right)\right)\right\|_2 \\ &\leq O(1)\left\|\mathbb{E}\left(K_h(X_i)\tilde{Z}_i\left(Y_i - \tilde{Z}_i^\top\gamma_n\right)\right)\right\|_2\end{aligned}$$

whereby Assumption 4 is used. Now set

$$f(x) = \mathbb{E}\left(\tilde{Z}_i\left(Y_i - \tilde{Z}_i^\top\gamma_n\right) \;\middle|\; X_i = x\right),$$

then

$$\mathbb{E}\left(K_h(X_i)\tilde{Z}_i\left(Y_i - \tilde{Z}_i^\top\gamma_n\right)\right) = \mathbb{E}\left(\frac{1}{h}K\left(\frac{X_i}{h}\right)f(X_i)\right).$$

We have

$$\begin{aligned}f_+ &= \mu_{\tilde{Z}Y+} - \mu_{\tilde{Z}\tilde{Z}^\top\gamma_n+} \\ &= \mu_{(Z-M_n^\top V)Y+} - \mu_{(Z-M_n^\top V)(Z-M_n^\top V)^\top+}\gamma_n \\ &= \mu_{ZY+} - \mu_{Z+}\mu_{Y+} + \mu_{Z+}\mu_{Y+} - \mu_{M_n^\top VY+} - (\mu_{ZZ^\top+} - \mu_{Z+}\mu_{Z^\top+})\gamma_n \\ &\quad + (-\mu_{Z+}\mu_{Z^\top+} + \mu_{M_n^\top VZ^\top+} + \mu_{ZV^\top M_n+} - \mu_{M_n^\top VV^\top M_n+})\gamma_n \\ &= \sigma^2_{ZY+} - \sigma^2_{Z+}\gamma_n \\ &\quad + \mu_{Z+}\mu_{Y+} - \mu_{M_n^\top VY+} \\ &\quad + (-\mu_{Z+}\mu_{Z^\top+} + \mu_{M_n^\top VZ^\top+} + \mu_{ZV^\top M_n+} - \mu_{M_n^\top VV^\top M_n+})\gamma_n \\ &= \sigma^2_{ZY+} - \sigma^2_{Z+}\gamma_n\end{aligned} \tag{7.8}$$

whereby the last equality holds due to

$$\begin{aligned}\mu_{M_n^\top VY+} &= \lim_{x\searrow 0}\mathbb{E}\left(\begin{pmatrix}\mu_Z(0) & 0 & h\mu'_{Z-} & h(\mu'_{Z+} - \mu'_{Z-})\end{pmatrix}\begin{pmatrix}1 \\ T_i \\ X_i/h \\ T_iX_i/h\end{pmatrix}\;\middle|\; X_i = x\right) \\ &= \lim_{x\searrow 0}\mathbb{E}\left((\mu_Z(0) + \mu'_{Z-}X_i + T_iX_i(\mu'_{Z+} - \mu'_{Z-}))Y_i \mid X_i = x\right) \\ &= \mu_Z(0)\mu_{Y+} \\ &= \mu_{Z+}\mu_{Y+},\end{aligned}$$

$$
\begin{aligned}
&\mu_{M_n^\top V V^\top M_n +} \\
&= \lim_{x \searrow 0} \mathbb{E}\left(\begin{pmatrix} \mu_Z(0) \\ 0 \\ h\mu'_{Z-} \\ h(\mu'_{Z+} - \mu'_{Z-}) \end{pmatrix}^\top \begin{pmatrix} 1 \\ T_i \\ X_i/h \\ T_i X_i/h \end{pmatrix} \begin{pmatrix} 1 \\ T_i \\ X_i/h \\ T_i X_i/h \end{pmatrix}^\top \begin{pmatrix} \mu_Z(0)^\top \\ 0 \\ h\mu'^\top_{Z-} \\ h(\mu'_{Z+} - \mu'_{Z-})^\top \end{pmatrix} \,\middle|\, X_i = x \right) \\
&= \lim_{x \searrow 0} \mathbb{E}\Big((\mu_Z(0) + \mu'_{Z-} X_i + T_i X_i(\mu'_{Z+} - \mu'_{Z-}))(\mu_Z(0)^\top + X_i \mu'^\top_{Z-} + X_i T_i (\mu'_{Z+} - \mu'_{Z-})^\top) \\
&\qquad\qquad \mid X_i = x \Big) \\
&= \mu_Z(0)\mu_Z(0)^\top \\
&= \mu_{Z+}\mu_{Z+}^\top,
\end{aligned}
$$

$$
\begin{aligned}
\mu_{M_n^\top V Z^\top +} &= \lim_{x \searrow 0} \mathbb{E}\left((\mu_Z(0) + \mu'_{Z-} X_i + T_i X_i(\mu'_{Z+} - \mu'_{Z-})) Z_i^\top \mid X_i = x \right) \\
&= \mu_Z(0)\mu_{Z+}^\top \\
&= \mu_{Z+}\mu_{Z+}^\top
\end{aligned}
$$

and finally

$$
\mu_{Z V^\top M_n +} = \left(\mu_{M_n^\top V Z^\top +}\right)^\top = \left(\mu_{Z+}\mu_{Z+}^\top\right)^\top = \mu_{Z+}\mu_{Z+}^\top.
$$

Analogously we obtain

$$
f_- = \sigma^2_{ZY-} - \sigma^2_{Z-}\gamma_n
$$

leading to

$$
\begin{aligned}
f_+ + f_- &= \sigma^2_{ZY+} - \sigma^2_{Z+}\gamma_n + \sigma^2_{ZY-} - \sigma^2_{Z-}\gamma_n \\
&= \sigma^2_{ZY+} + \sigma^2_{ZY-} - (\sigma^2_{Z+} + \sigma^2_{Z-})\gamma_n \\
&= \sigma^2_{ZY+} + \sigma^2_{ZY-} - (\sigma^2_{Z+} + \sigma^2_{Z-})\left((\sigma^2_{Z+} + \sigma^2_{Z-})^{-1}(\sigma^2_{ZY+} + \sigma^2_{ZY-})\right) \\
&= 0.
\end{aligned}
$$

We perform a Taylor expansion similar as in (5.2), but this time just up to the first degree, which means

$$
\begin{aligned}
\mathbb{E}\left(K_h(X_i)\tilde{Z}_i \left(Y_i - \tilde{Z}_i^\top \gamma_n \right) \right) &= K_-^{(0)} f_- f_X(0) + K_+^{(0)} f_+ f_X(0) + O(h) \\
&= K_-^{(0)} f_X(0) \underbrace{(f_- + f_+)}_{=0} + O(h) = O(h).
\end{aligned}
$$

Hence there is some constant $C > 0$ such that

$$
\left\| \check{\beta}_n - \gamma_n \right\|_2 \leq Ch \xrightarrow{n \to \infty} 0,
$$

which in turn implies the statement of the lemma. □

Lemma 7.2 (Computation for the variance)**.** *Let the assumptions of Section 7.1 hold. Then*

$$\frac{1}{h}\mathbb{E}\left(K\left(\frac{X_i}{h}\right)^2 (w^\top V_i)^2 \check{r}_i(h)^2\right) + o(1) = \mathcal{S}_n^2$$

whereby $w = \left(\left[(f_X(0)\kappa(K))^{-1}\right]_{2\cdot}\right)^\top$.

Proof. We conclude in two steps. First, we show that

$$\frac{1}{h}\mathbb{E}\left(K\left(\frac{X_i}{h}\right)^2 (w^\top V_i)^2 \check{r}_i(h)^2\right) \xrightarrow{h\to 0} \frac{\mathcal{C}_\mathcal{S}}{f_X(0)}\left(\sigma_l^2 + \sigma_r^2\right), \tag{7.9}$$

whereby σ_l^2 and σ_r^2 are the finite numbers provided by Assumption 7. Indeed,

$$\begin{aligned}
&\frac{1}{h}\mathbb{E}\left(K\left(\frac{X_i}{h}\right)^2 (w^\top V_i)^2 \check{r}_i(h)^2\right) \\
&= \frac{1}{h}\mathbb{E}\left(K\left(\frac{X_i}{h}\right)^2 (w^\top V_i)^2 \mathbb{E}\left(\check{r}_i(h)^2 \mid X_i\right)\right) \\
&= \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{x}{h}\right)^2 \left(w^\top \begin{pmatrix} 1 & \mathbb{1}(x \geq 0) & \frac{x}{h} & \frac{x}{h}\mathbb{1}(x \geq 0)\end{pmatrix}^\top\right)^2 \mu_{\check{r}_i(h)^2}(x) f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} K(y)^2 \frac{1}{f_X(0)^2}\left([\kappa(K)^{-1}]_2 \begin{pmatrix} 1 & \mathbb{1}(yh \geq 0) & y & y\mathbb{1}(yh \geq 0)\end{pmatrix}^\top\right)^2 \\
&\qquad \mu_{\check{r}_i(h)^2}(yh) f_X(yh)\, dy \\
&= \int_{-1}^{0} K(y)^2 \frac{1}{f_X(0)^2}\left([\kappa(K)^{-1}]_2 \begin{pmatrix} 1 & 0 & y & 0\end{pmatrix}^\top\right)^2 \mu_{\check{r}_i(h)^2}(yh) f_X(yh)\, dy \\
&\quad + \int_{0}^{1} K(y)^2 \frac{1}{f_X(0)^2}\left([\kappa(K)^{-1}]_2 \begin{pmatrix} 1 & 1 & y & y\end{pmatrix}^\top\right)^2 \mu_{\check{r}_i(h)^2}(yh) f_X(yh)\, dy \\
&\xrightarrow{h\to 0} \int_{-1}^{0} K(y)^2 \frac{1}{f_X(0)}\left([\kappa(K)^{-1}]_2 \begin{pmatrix} 1 & 0 & y & 0\end{pmatrix}^\top\right)^2 \sigma_l^2\, dy \\
&\quad + \int_{0}^{1} K(y)^2 \frac{1}{f_X(0)}\left([\kappa(K)^{-1}]_2 \begin{pmatrix} 1 & 1 & y & y\end{pmatrix}^\top\right)^2 \sigma_r^2\, dy
\end{aligned}$$

whereby the limit is moved inside the integral by the Theorem of Dominated Convergence. Also, we used that

$$\mu_{\check{r}_i(h)^2}(-yh) \xrightarrow[n\to\infty]{} \sigma_l^2 \quad \text{and} \quad \mu_{\check{r}_i(h)^2}(yh) \xrightarrow[n\to\infty]{} \sigma_r^2$$

for all $y \in [0,1]$ since $\check{r}_i(h) = r_i(h)$ and by the choice of σ_l and σ_r in Assumption 7.

Next, we use the representation of $\kappa(K)^{-1}$ as stated in Lemma 5.3 and obtain

$$\begin{aligned}
&\int_{-\infty}^{0} K(y)^2 \frac{1}{f_X(0)} \left([\kappa(K)^{-1}]_2 \begin{pmatrix} 1 & 0 & y & 0 \end{pmatrix}^\top \right)^2 \sigma_l^2 \, dy \\
&= \frac{f_X(0)^{-1}}{\left[\left(K_+^{(1)}\right)^2 - \frac{1}{2} K_+^{(2)} \right]^2} \int_{-\infty}^{0} K(y)^2 \left(\begin{pmatrix} K_+^{(2)} & -2K_+^{(2)} & K_+^{(1)} & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & y & 0 \end{pmatrix}^\top \right)^2 \sigma_l^2 \, dy \\
&= \frac{f_X(0)^{-1}}{\left[\left(K_+^{(1)}\right)^2 - \frac{1}{2} K_+^{(2)} \right]^2} \int_{-\infty}^{0} K(y)^2 \left(\left(K_+^{(2)}\right)^2 + y^2 \left(K_+^{(1)}\right)^2 + 2y K_+^{(2)} K_+^{(1)} \right) \sigma_l^2 \, dy \\
&= \frac{f_X(0)^{-1}}{\left[\left(K_+^{(1)}\right)^2 - \frac{1}{2} K_+^{(2)} \right]^2} \left((K^2)_-^{(0)} \left(K_+^{(2)}\right)^2 + (K^2)_-^{(2)} \left(K_+^{(1)}\right)^2 + 2(K^2)_-^{(1)} K_+^{(2)} K_+^{(1)} \right) \sigma_l^2 \\
&= \frac{f_X(0)^{-1}}{\left[\left(K_+^{(1)}\right)^2 - \frac{1}{2} K_+^{(2)} \right]^2} \left((K^2)_+^{(0)} \left(K_+^{(2)}\right)^2 + (K^2)_+^{(2)} \left(K_+^{(1)}\right)^2 - 2(K^2)_+^{(1)} K_+^{(2)} K_+^{(1)} \right) \sigma_l^2 \\
&= \frac{\mathcal{C}_{\mathcal{S}}}{f_X(0)} \sigma_l^2
\end{aligned}$$

and

$$\begin{aligned}
&\int_{0}^{\infty} K(y)^2 \frac{1}{f_X(0)} \left([\kappa(K)^{-1}]_2 \begin{pmatrix} 1 & 1 & y & y \end{pmatrix}^\top \right)^2 \sigma_r^2 \, dy \\
&= \frac{f_X(0)^{-1}}{\left[\left(K_+^{(1)}\right)^2 - \frac{1}{2} K_+^{(2)} \right]^2} \int_{0}^{\infty} K(y)^2 \left(\begin{pmatrix} K_+^{(2)} & -2K_+^{(2)} & K_+^{(1)} & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 & y & y \end{pmatrix}^\top \right)^2 \sigma_r^2 \, dy \\
&= \frac{f_X(0)^{-1}}{\left[\left(K_+^{(1)}\right)^2 - \frac{1}{2} K_+^{(2)} \right]^2} \int_{0}^{\infty} K(y)^2 \left(\left(K_+^{(2)}\right)^2 + y^2 \left(K_+^{(1)}\right)^2 - 2y K_+^{(2)} K_+^{(1)} \right) \sigma_r^2 \, dy \\
&= \frac{f_X(0)^{-1}}{\left[\left(K_+^{(1)}\right)^2 - \frac{1}{2} K_+^{(2)} \right]^2} \left((K^2)_+^{(0)} \left(K_+^{(2)}\right)^2 + (K^2)_+^{(2)} \left(K_+^{(1)}\right)^2 - 2(K^2)_+^{(1)} K_+^{(2)} K_+^{(1)} \right) \sigma_r^2 \\
&= \frac{\mathcal{C}_{\mathcal{S}}}{f_X(0)} \sigma_r^2.
\end{aligned}$$

Therefore, the convergence in (7.9) is proved. We proceed with the second step and want to show that

$$\sigma_{\tilde{Y}+}^2 = \sigma_r^2 \quad \text{and} \quad \sigma_{\tilde{Y}-}^2 = \sigma_l^2. \tag{7.10}$$

We do this by first using the representation of $\check{\theta}_0(h)$ as stated in (5.10). We have

$$\begin{aligned}
\check{\theta}_0(h) &= \Big(I - \mathbb{E}\left(K_h(X_i)V_iV_i^\top\right)^{-1}\mathbb{E}\left(K_h(X_i)V_i\tilde{Z}_i^\top\right)\mathbb{E}\left(K_h(X_i)\tilde{Z}_i\tilde{Z}_i^\top\right)^{-1} \\
&\quad \mathbb{E}\left(K_h(X_i)\tilde{Z}_iV_i^\top\right)\Big)^{-1}\left(\kappa(K)^{-1}\mathbb{E}\left(K_h(X_i)V_iV_i^\top\right)\right)^{-1} \\
&\quad \kappa(K)^{-1}\Big(\mathbb{E}(K_h(X_i)V_iY_i) - \mathbb{E}\left(K_h(X_i)V_i\tilde{Z}_i^\top\right)\mathbb{E}\left(K_h(X_i)\tilde{Z}_i\tilde{Z}_i^\top\right)^{-1} \\
&\quad \mathbb{E}(K_h(X_i)\tilde{Z}_iY_i)\Big) \\
&= \left(I - \left(\frac{1}{f_X(0)}\kappa(K)^{-1} + o(1)\right)\cdot o(1)\right)^{-1}\cdot\left(\frac{1}{f_X(0)}I + o(1)\right) \\
&\quad \cdot\left(\begin{pmatrix} f_X(0)\mu_{Y-} \\ f_X(0)\tau_Y \\ h[\mu_Y f_X]'_- \\ h\left([\mu_Y f_X]'_+ - [\mu_Y f_X]'_-\right)\end{pmatrix} + o(1)\right) \\
&= \begin{pmatrix}\mu_{Y-}\\ \tau_Y \\ 0 \\ 0\end{pmatrix} + o(1) \quad \xrightarrow{h\to 0} \quad \begin{pmatrix}\mu_{Y-}\\ \tau_Y \\ 0 \\ 0\end{pmatrix}
\end{aligned} \tag{7.11}$$

whereby we used Lemma 5.2, Lemma 5.4 and Lemma 5.5 together with the fact that $\mu_{\tilde{Z}}(0) = \mu'_{\tilde{Z}}(0) = 0$ to obtain the above convergence.

Next, we want to state that

$$\gamma_0(h) = \gamma_n + O(h). \tag{7.12}$$

To show this we use the representation given in (5.10) again, but this time we interchange the roles of V_i and Z_i in order to represent $\gamma_0(h)$ instead of $\theta_0(h)$. Indeed, we obtain

$$\begin{aligned}
&\gamma_0(h) \\
&= \Big(\mathbb{E}(K_h(X_i)Z_iZ_i^\top) - \mathbb{E}(K_h(X_i)Z_iV_i^\top)\mathbb{E}(K_h(X_i)V_iV_i^\top)^{-1}\mathbb{E}(K_h(X_i)V_iZ_i^\top)\Big)^{-1} \\
&\quad \Big(\mathbb{E}(K_h(X_i)Z_iY_i) - \mathbb{E}(K_h(X_i)Z_iV_i^\top)\mathbb{E}(K_h(X_i)V_iV_i^\top)^{-1}\mathbb{E}(K_h(X_i)V_iY_i)\Big) \\
&= \Big(\mathbb{E}(K_h(X_i)Z_iZ_i^\top) - \mathbb{E}(K_h(X_i)Z_iV_i^\top)\left(\kappa(K)^{-1}\mathbb{E}(K_h(X_i)V_iV_i^\top)\right)^{-1} \\
&\quad \kappa(K)^{-1}\mathbb{E}(K_h(X_i)V_iZ_i^\top)\Big)^{-1} \\
&\quad \Big(\mathbb{E}(K_h(X_i)Z_iY_i) - \mathbb{E}(K_h(X_i)Z_iV_i^\top)\left(\kappa(K)^{-1}\mathbb{E}(K_h(X_i)V_iV_i^\top)\right)^{-1} \\
&\quad \kappa(K)^{-1}\mathbb{E}(K_h(X_i)V_iY_i)\Big).
\end{aligned} \tag{7.13}$$

We begin by examining the first factor. We have

$$\begin{aligned}\mathbb{E}(K_h(X_i)Z_iZ_i^\top) &= \mathbb{E}(K_h(X_i)\mathbb{E}(Z_iZ_i^\top \mid X_i)) \\ &= K_-^{(0)}\mu_{ZZ^\top -}f_X(0) + K_+^{(0)}\mu_{ZZ^\top +}f_X(0) \\ &= \frac{f_X(0)}{2}\left(\mu_{ZZ^\top +} + \mu_{ZZ^\top -}\right)\end{aligned}$$

by performing Taylor expansion as in (5.2),

$$\begin{aligned}\mathbb{E}(K_h(X_i)Z_iV_i^\top) &= \mathbb{E}\left(K_h(X_i)Z_i\begin{pmatrix}1 & T_i & \frac{X_i}{h} & \frac{T_iX_i}{h}\end{pmatrix}\right) \\ &= \begin{pmatrix}\mathbb{E}(K_h(X_i)\mathbb{E}(Z_i \mid X_i)) & * & * & *\end{pmatrix} \\ &= \begin{pmatrix}\frac{f_X(0)}{2}\left(\mu_{Z+} + \mu_{Z-}\right) & * & * & *\end{pmatrix}\end{aligned}$$

also by Taylor expansion according to (5.2),

$$\kappa(K)^{-1}\mathbb{E}(K_h(X_i)V_iV_i^\top) = f_X(0)I + O(h)$$

by Lemma 5.4, and finally

$$\kappa(K)^{-1}\mathbb{E}(K_h(X_i)V_iZ_i^\top) = \begin{pmatrix}f_X(0)\mu_{Z-}^\top \\ 0 \\ 0 \\ 0\end{pmatrix} + O(h)$$

by applying Lemma 5.2 compontent-wise and the fact that $\mu_{Z+} = \mu_{Z-}$ due to the continuity of μ_Z. Taking all the above statements together gives

$$\begin{aligned}&\mathbb{E}(K_h(X_i)Z_iZ_i^\top) - \mathbb{E}(K_h(X_i)Z_iV_i^\top)\left(\kappa(K)^{-1}\mathbb{E}(K_h(X_i)V_iV_i^\top)\right)^{-1} \\ &\kappa(K)^{-1}\mathbb{E}(K_h(X_i)V_iZ_i^\top) \\ &= \frac{f_X(0)}{2}\left(\mu_{ZZ^\top +} + \mu_{ZZ^\top -}\right) - \begin{pmatrix}\frac{f_X(0)}{2}\left(\mu_{Z+} + \mu_{Z-}\right) & * & * & *\end{pmatrix}\frac{1}{f_X(0)}\begin{pmatrix}f_X(0)\mu_{Z-}^\top \\ 0 \\ 0 \\ 0\end{pmatrix} \\ &\quad + O(h) \\ &= \frac{f_X(0)}{2}\left(\mu_{ZZ^\top +} + \mu_{ZZ^\top -} - \mu_{Z+}\mu_{Z+}^\top - \mu_{Z-}\mu_{Z-}^\top\right) + O(h) \\ &= \frac{f_X(0)}{2}\left(\sigma_{Z+}^2 + \sigma_{Z-}^2\right) + O(h)\end{aligned}$$

whereby we used again that $\mu_{Z+}^\top = \mu_{Z-}^\top$. In an analogous way, we obtain for the second

factor

$$\begin{aligned}&\mathbb{E}(K_h(X_i)Z_iY_i) - \mathbb{E}(K_h(X_i)Z_iV_i^\top)\left(\kappa(K)^{-1}\mathbb{E}(K_h(X_i)V_iV_i^\top)\right)^{-1}\kappa(K)^{-1}\mathbb{E}(K_h(X_i)V_iY_i)\\ &= \frac{f_X(0)}{2}\left(\sigma^2_{ZY+} + \sigma^2_{ZY-}\right) + O(h).\end{aligned}$$

Refering back to equation (7.13) and inserting the respective terms, we overall obtain

$$\begin{aligned}\gamma_0(h) &= \left(\frac{f_X(0)}{2}\left(\sigma^2_{Z+} + \sigma^2_{Z-}\right) + O(h)\right)^{-1}\left(\frac{f_X(0)}{2}\left(\sigma^2_{ZY+} + \sigma^2_{ZY-}\right) + O(h)\right)\\ &= \left(\sigma^2_{Z+} + \sigma^2_{Z-}\right)^{-1}\left(\sigma^2_{ZY+} + \sigma^2_{ZY-}\right) + O(h)\\ &= \gamma_n + O(h).\end{aligned} \tag{7.14}$$

We can use this equality in order to get

$$\begin{aligned}&\theta_0(h) - \begin{pmatrix}\mu_{\tilde{Y}-}\\ \mu_{\tilde{Y}+} - \mu_{\tilde{Y}-}\\ 0\\ 0\end{pmatrix}\\ &= \check{\theta}_0(h) - M_n\gamma_0(h) - \begin{pmatrix}\mu_{Y-} - \mu_{Z-}^\top\gamma_n\\ \tau_Y\\ 0\\ 0\end{pmatrix}\\ &= \check{\theta}_0(h) - \begin{pmatrix}\mu_Z(0)^\top\\ 0\\ h\mu_{Z-}'^\top\\ h(\mu_{Z+}'^\top - \mu_{Z-}'^\top)\end{pmatrix}(\gamma_n + O(h)) - \begin{pmatrix}\mu_{Y-} - \mu_{Z-}^\top\gamma_n\\ \tau_Y\\ 0\\ 0\end{pmatrix}\\ &= \check{\theta}_0(h) - \begin{pmatrix}\mu_Z(0)^\top(\gamma_n + O(h))\\ 0\\ h\mu_{Z-}'^\top(\gamma_n + O(h))\\ h(\mu_{Z+}'^\top - \mu_{Z-}'^\top)(\gamma_n + O(h))\end{pmatrix} - \begin{pmatrix}\mu_{Y-} - \mu_{Z-}^\top\gamma_n\\ \tau_Y\\ 0\\ 0\end{pmatrix}\\ &\xrightarrow[h\to 0]{(7.11)} \begin{pmatrix}\mu_{Y-}\\ \tau_Y\\ 0\\ 0\end{pmatrix} - \begin{pmatrix}\mu_{Z-}^\top\gamma_n\\ 0\\ 0\\ 0\end{pmatrix} - \begin{pmatrix}\mu_{Y-} - \mu_{Z-}^\top\gamma_n\\ \tau_Y\\ 0\\ 0\end{pmatrix} = 0.\end{aligned} \tag{7.15}$$

Next, we have

$$
\begin{aligned}
&\lim_{x\searrow 0}\mathbb{E}(r_i^2(h)\mid X_i=x)-\lim_{x\searrow 0}\mathrm{Var}(\tilde{Y}_i\mid X_i=x)\\
&=\lim_{x\searrow 0}\Big(\mathbb{E}\left((Y_i-V_i^\top\theta_0(h)-Z_i^\top\gamma_0(h))^2\mid X_i=x\right)-\mathbb{E}\left((Y_i-Z_i^\top\gamma_n)^2\mid X_i=x\right)\\
&\quad+\mathbb{E}\left(Y_i-Z_i^\top\gamma_n\mid X_i=x\right)^2\Big)\\
&=\lim_{x\searrow 0}\Big(\mathbb{E}\left(Y_i^2\mid X=x\right)-2\mathbb{E}\left(Y_iV_i^\top\theta_0(h)\mid X_i=x\right)-2\mathbb{E}\left(Y_iZ_i^\top\gamma_0(h)\mid X_i=x\right)\\
&\quad+\mathbb{E}\left((V_i^\top\theta_0(h))^2\mid X_i=x\right)+2\mathbb{E}\left(V_i^\top\theta_0(h)Z_i^\top\gamma_0(h)\mid X_i=x\right)\\
&\quad+2\mathbb{E}\left((Z_i^\top\gamma_0(h))^2\mid X_i=x\right)-\mathbb{E}\left(Y_i^2\mid X_i=x\right)+\mathbb{E}\left(Y_i\mid X_i=x\right)^2\\
&\quad-\mathbb{E}\left((Z_i^\top\gamma_n)^2\mid X_i=x\right)+\mathbb{E}\left(Z_i^\top\gamma_n\mid X_i=x\right)^2+2\mathbb{E}\left(Y_iZ_i^\top\gamma_n\mid X_i=x\right)\\
&\quad-2\mathbb{E}\left(Y_i\mid X_i=x\right)\mathbb{E}\left(Z_i^\top\gamma_n\mid X_i=x\right)\Big)\\
&=\theta_0(h)^\top\mu_{VV^\top+}\theta_0(h)-\left(\mu_{Y+}-\mu_{Z+}^\top\gamma_n\right)^2\\
&\quad+\gamma_0(h)^\top\mu_{ZZ^\top+}\gamma_0(h)-\gamma_n^\top\mu_{ZZ^\top+}\gamma_n\\
&\quad-2\mu_{YV^\top+}\theta_0(h)+2\mu_{Y+}^2-2\mu_{Z+}^\top\mu_{Y+}\gamma_n\\
&\quad-2\mu_{YZ^\top+}\gamma_0(h)+2\mu_{YZ^\top+}^\top\gamma_n\\
&\quad+2\theta_0(h)^\top\mu_{VZ^\top+}\gamma_0(h)+2\gamma_n^\top\mu_{Z+}\mu_{Z+}^\top\gamma_n-2\mu_{Y+}\mu_{Z+}^\top\gamma_n.
\end{aligned}
\tag{7.16}
$$

We examine this expression row-wise for $h\to 0$ by using (7.14) as well as (7.15), and state that each of them converges to zero, leading to the whole expression converging to zero. Indeed,

$$
\begin{aligned}
&\theta_0(h)^\top\mu_{VV^\top+}\theta_0(h)\\
&\overset{(7.15)}{=}\begin{pmatrix}\mu_{\tilde{Y}-}\\ \mu_{\tilde{Y}+}-\mu_{\tilde{Y}-}\\ 0\\ 0\end{pmatrix}^\top\mu_{VV^\top+}\begin{pmatrix}\mu_{\tilde{Y}-}\\ \mu_{\tilde{Y}+}-\mu_{\tilde{Y}-}\\ 0\\ 0\end{pmatrix}+o(1)\\
&=\begin{pmatrix}\mu_{\tilde{Y}-}\\ \mu_{\tilde{Y}+}-\mu_{\tilde{Y}-}\\ 0\\ 0\end{pmatrix}^\top\lim_{x\searrow 0}\mathbb{E}\left(\begin{pmatrix}1 & T_i & \frac{X_i}{h} & \frac{T_iX_i}{h}\\ T_i & T_i^2 & \frac{T_iX_i}{h} & \frac{T_i^2X_i}{h}\\ \frac{X_i}{h} & \frac{T_iX_i}{h} & \frac{X_i^2}{h^2} & \frac{T_iX_i^2}{h^2}\\ \frac{T_iX_i}{h} & \frac{T_i^2X_i}{h} & \frac{T_iX_i^2}{h^2} & \frac{T_i^2X_i^2}{h^2}\end{pmatrix}\,\middle|\, X_i=x\right)\begin{pmatrix}\mu_{\tilde{Y}-}\\ \mu_{\tilde{Y}+}-\mu_{\tilde{Y}-}\\ 0\\ 0\end{pmatrix}\\
&\quad+o(1)\\
&=\begin{pmatrix}\mu_{\tilde{Y}-} & \mu_{\tilde{Y}+}-\mu_{\tilde{Y}-} & 0 & 0\end{pmatrix}\begin{pmatrix}1 & 1 & 0 & 0\\ 1 & 1 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{pmatrix}\begin{pmatrix}\mu_{\tilde{Y}-}\\ \mu_{\tilde{Y}+}-\mu_{\tilde{Y}-}\\ 0\\ 0\end{pmatrix}+o(1)
\end{aligned}
$$

$$
\begin{aligned}
&= \begin{pmatrix} \mu_{\tilde{Y}-} & \mu_{\tilde{Y}+} - \mu_{\tilde{Y}-} & 0 & 0 \end{pmatrix} \begin{pmatrix} \mu_{\tilde{Y}+} \\ \mu_{\tilde{Y}+} \\ 0 \\ 0 \end{pmatrix} + o(1) \\
&= \mu_{\tilde{Y}+}^2 + o(1) \\
&= \left(\mu_{Y+} - \mu_{Z^\top +}\gamma_n\right)^2 + o(1),
\end{aligned}
$$

$$
\gamma_0(h)^\top \mu_{ZZ^\top +}\gamma_0(h) - \gamma_n^\top \mu_{ZZ^\top +}\gamma_n \overset{(7.14)}{=} O(h),
$$

$$
\begin{aligned}
\mu_{YV^\top +}\theta_0 &\overset{(7.15)}{=} \lim_{x \searrow 0} \mathbb{E}\left(\begin{pmatrix} Y_i & Y_i T_i & \frac{Y_i X_i}{h} & \frac{Y_i T_i X_i}{h} \end{pmatrix} \mid X_i = x \right) \begin{pmatrix} \mu_{\tilde{Y}-} \\ \mu_{\tilde{Y}+} - \mu_{\tilde{Y}-} \\ 0 \\ 0 \end{pmatrix} + o(1) \\
&= \begin{pmatrix} \mu_{Y+} & \mu_{Y+} & 0 & 0 \end{pmatrix} \begin{pmatrix} \mu_{\tilde{Y}-} \\ \mu_{\tilde{Y}+} - \mu_{\tilde{Y}-} \\ 0 \\ 0 \end{pmatrix} + o(1) \\
&= \mu_{Y+}(\mu_{Y-} - \mu_{Z^\top -}\gamma_n) + \mu_{Y+}(\mu_{Y+} - \mu_{Y-} + \mu_{Z^\top -}\gamma_n - \mu_{Z^\top +}\gamma_n) + o(1) \\
&= \mu_{Y+}^2 - \mu_{Y+}\mu_{Z^\top +}\gamma_n + o(1),
\end{aligned}
$$

$$
\mu_{YZ^\top +}\gamma_0 - \mu_{YZ^\top +}\gamma_n \overset{(7.14)}{=} O(h),
$$

and at last we have

$$
\begin{aligned}
&\theta_0(h)^\top \mu_{VZ^\top +}\gamma_0 \\
&\overset{(7.14)}{=} \theta_0(h)^\top \mu_{VZ^\top +}\gamma_n + \theta_0(h)^\top \mu_{VZ^\top +} O(h) \\
&\overset{(7.15)}{=} \begin{pmatrix} \mu_{\tilde{Y}-} & \mu_{\tilde{Y}+} - \mu_{\tilde{Y}-} & 0 & 0 \end{pmatrix} \lim_{x \searrow 0} \left(\begin{pmatrix} 1 \\ \mathbb{1}(x \geq 0) \\ x/h \\ \mathbb{1}(x \geq 0)x/h \end{pmatrix} \mathbb{E}(Z_i^\top \mid X_i = x) \right) \gamma_n + o(1) \\
&= \begin{pmatrix} \mu_{\tilde{Y}-} & \mu_{\tilde{Y}+} - \mu_{\tilde{Y}-} & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} \mu_{Z^\top +}\gamma_n + o(1) \\
&= \mu_{\tilde{Y}+}\mu_{Z^\top +}\gamma_n + o(1) = \mu_{Y+}\mu_{Z^\top +}\gamma_n - \gamma_n^\top \mu_{Z+}\mu_{Z+}^\top \gamma_n + o(1).
\end{aligned}
$$

Overall, we can merge these statements and conclude by refering back to equation (7.16)

that

$$\lim_{x \searrow 0} \mathbb{E}(r_i^2(h) \mid X_i = x) \xrightarrow[n\to\infty]{} \lim_{x \searrow 0} \operatorname{Var}\left(\tilde{Y}_i \mid X_i = x\right).$$

Completely analogously, we also obtain

$$\lim_{x \nearrow 0} \mathbb{E}(r_i^2(h) \mid X_i = x) \xrightarrow[n\to\infty]{} \lim_{x \nearrow 0} \operatorname{Var}\left(\tilde{Y}_i \mid X_i = x\right).$$

By definition of σ_l and σ_r, we can conclude that

$$\sigma^2_{\tilde{Y}+} = \sigma_r^2 \quad \text{and} \quad \sigma^2_{\tilde{Y}-} = \sigma_l^2,$$

which is exactly statement (7.10) that we wanted to show. Hence, the whole assertion is proved. □

7.3 Consequences

As a consequence of the representation of the bias and the variance, we want to discuss if additional covariates actually enhance our estimator and its MSE. Therefore, as done in [KR21], we want to study the bias and variance in comparison to those of the baseline estimator without covariates.

7.3.1 Discussion of Bias

The first term of the bias representation can be written as

$$\mu''_{\tilde{Y}+} - \mu''_{\tilde{Y}-} = \left(\mu''_{Y+} - \mu''_{Y-}\right) - \left(\mu''_{Z+} - \mu''_{Z-}\right)^\top \gamma_n.$$

We can see that the first summand equals the bias of the baseline estimator in (3.1) up to a constant, whereby the second summand describes the impact of the covariates on the bias. So in order to understand whether the bias is better than in the case of the baseline estimator, we have to examine the second summand. Indeed, the covariates can contribute to the bias in both a favorable and unfavorable way. The best case would be that the second term makes the bias smaller and hence decreases the error of our estimator. But of course, the term can also vanish or even make the bias larger. Nonetheless, as argued in [KR21], we do not have to be afraid of that too much. This is due to the fact that we deal with pre-treatment covariates, which means that the treatment has no impact on them. This is expressed by the continuity assumption imposed on μ_Z leading to $\mu_{Z+} = \mu_{Z-}$, i.e. μ_Z not jumping at the cutoff. Now, as we could argue that $\mu''_{Z+} = \mu''_{Z-}$ is just a slightly stronger version of that assumption, we could assume that $\mu''_{Z+} - \mu''_{Z-} = 0$. Then, the second term would vanish and we would obtain the bias of the baseline estimator.

7.3.2 Discussion of Variance

We want to show that the variance of the covariate-adjusted estimator is less or equal, but does not exceed the variance of the baseline estimator under the condition that the left- and right-sided limit of the covariance matrix of Z_i are invertible, i.e. we need the additional assumption that

$$\lim_{x \searrow 0} \left(\mathrm{Cov}\left(Z_i^{(k)}, Z_i^{(l)} \mid X_i = x \right) \right)_{k,l \in \{1,\dots,p\}} \text{ and } \lim_{x \nearrow 0} \left(\mathrm{Cov}\left(Z_i^{(k)}, Z_i^{(l)} \mid X_i = x \right) \right)_{k,l \in \{1,\dots,p\}}$$

are invertible. With this condition being satisfied, the idea of the proof is to show that γ_n minimizes the function

$$\gamma \mapsto \lim_{x \searrow 0} \mathrm{Var}\left(Y_i - Z_i^\top \gamma \mid X_i = x \right) + \lim_{x \nearrow 0} \mathrm{Var}\left(Y_i - Z_i^\top \gamma \mid X_i = x \right)$$

leading to

$$\sigma^2_{\tilde{Y}_+} + \sigma^2_{\tilde{Y}_-} \leq \sigma^2_{Y_+} + \sigma^2_{Y_-}$$

whereby the expression on the right-hand side is the leading term of the baseline estimator's variance provided in (3.2). Next, we want to discuss this idea in more detail.

Proof (that γ_n minimizes the function). We have for $k \in \{1, ..., p\}$

$$\begin{aligned}
&\frac{\partial}{\partial \gamma^k} \left(\lim_{x \searrow 0} \mathrm{Var}\left(Y_i - Z_i^\top \gamma \mid X_i = x \right) + \lim_{x \nearrow 0} \mathrm{Var}\left(Y_i - Z_i^\top \gamma \mid X_i = x \right) \right) \\
&= \frac{\partial}{\partial \gamma^k} \Bigg(\lim_{x \searrow 0} \mathbb{E}\left(\left(Y_i - Z_i^\top \gamma \right)^2 \mid X_i = x \right) - \lim_{x \searrow 0} \mathbb{E}\left(Y_i - Z_i^\top \gamma \mid X_i = x \right)^2 \\
&\quad + \lim_{x \nearrow 0} \mathbb{E}\left(\left(Y_i - Z_i^\top \gamma \right)^2 \mid X_i = x \right) - \lim_{x \nearrow 0} \mathbb{E}\left(Y_i - Z_i^\top \gamma \mid X_i = x \right)^2 \Bigg) \\
&= \lim_{x \searrow 0} \mathbb{E}\left(2 Z_i^{(k)} \left(Y_i - Z_i^\top \gamma \right) \mid X_i = x \right) - \lim_{x \searrow 0} 2\mathbb{E}\left(Z_i^{(k)} \mid X_i = x \right) \mathbb{E}\left(Y_i - Z_i^\top \gamma \mid X_i = x \right) \\
&\quad + \lim_{x \nearrow 0} \mathbb{E}\left(2 Z_i^{(k)} \left(Y_i - Z_i^\top \gamma \right) \mid X_i = x \right) - \lim_{x \nearrow 0} 2\mathbb{E}\left(Z_i^{(k)} \mid X_i = x \right) \mathbb{E}\left(Y_i - Z_i^\top \gamma \mid X_i = x \right)
\end{aligned}$$

hence

$$\begin{aligned}
&\nabla_\gamma \left(\lim_{x \searrow 0} \mathrm{Var}\left(Y_i - Z_i^\top \gamma \mid X_i = x \right) + \lim_{x \nearrow 0} \mathrm{Var}\left(Y_i - Z_i^\top \gamma \mid X_i = x \right) \right) \\
&= \lim_{x \searrow 0} \mathbb{E}\left(2 Z_i \left(Y_i - Z_i^\top \gamma \right) \mid X_i = x \right) - \lim_{x \searrow 0} 2\mathbb{E}\left(Z_i \mid X_i = x \right) \mathbb{E}\left(Y_i - Z_i^\top \gamma \mid X_i = x \right) \\
&\quad + \lim_{x \nearrow 0} \mathbb{E}\left(2 Z_i \left(Y_i - Z_i^\top \gamma \right) \mid X_i = x \right) - \lim_{x \nearrow 0} 2\mathbb{E}\left(Z_i \mid X_i = x \right) \mathbb{E}\left(Y_i - Z_i^\top \gamma \mid X_i = x \right) \\
&= 2\left(\mu_{ZY+} - \mu_{ZZ^\top \gamma +} - \mu_{Z+}\left(\mu_{Y+} - \mu_{Z^\top \gamma +} \right) + \mu_{ZY-} - \mu_{ZZ^\top \gamma -} - \mu_{Z-}\left(\mu_{Y-} - \mu_{Z^\top \gamma -} \right) \right) \\
&= 2\left(\sigma^2_{ZY-} + \sigma^2_{ZY+} + \mu_{Z+}\mu_{Z^\top +}\gamma - \mu_{ZZ^\top +}\gamma + \mu_{Z-}\mu_{Z^\top -}\gamma - \mu_{ZZ^\top -}\gamma \right) \\
&= 2\left(\sigma^2_{ZY-} + \sigma^2_{ZY+} - \left(\sigma^2_{Z-} + \sigma^2_{Z+} \right)\gamma \right).
\end{aligned}$$

Evaluating the gradient at γ_n gives

$$\begin{aligned}
&\nabla_{\gamma=\gamma_n}\left(\lim_{x\searrow 0}\operatorname{Var}\left(Y_i - Z_i^\top\gamma \mid X_i = x\right) + \lim_{x\nearrow 0}\operatorname{Var}\left(Y_i - Z_i^\top\gamma \mid X_i = x\right)\right)\\
&= 2\left(\sigma^2_{ZY-} + \sigma^2_{ZY+} - \left(\sigma^2_{Z-} + \sigma^2_{Z+}\right)\left(\sigma^2_{Z-} + \sigma^2_{Z+}\right)^{-1}\left(\sigma^2_{ZY-} + \sigma^2_{ZY+}\right)\right)\\
&= 0.
\end{aligned}$$

It remains to show that the Hessian matrix is positive definite. Indeed for $k, l \in \{1, ..., p\}$

$$\begin{aligned}
&\frac{\partial^2}{\partial\gamma^l\partial\gamma^k}\left(\lim_{x\searrow 0}\operatorname{Var}\left(Y_i - Z_i^\top\gamma \mid X_i = x\right) + \lim_{x\nearrow 0}\operatorname{Var}\left(Y_i - Z_i^\top\gamma \mid X_i = x\right)\right)\\
&= \frac{\partial}{\partial\gamma^l}\Bigg(\lim_{x\searrow 0}\mathbb{E}\left(2Z_i^{(k)}\left(Y_i - Z_i^\top\gamma\right) \mid X_i = x\right)\\
&\quad - \lim_{x\searrow 0} 2\mathbb{E}\left(Z_i^{(k)} \mid X_i = x\right)\mathbb{E}\left(Y_i - Z_i^\top\gamma \mid X_i = x\right)\\
&\quad + \lim_{x\nearrow 0}\mathbb{E}\left(2Z_i^{(k)}\left(Y_i - Z_i^\top\gamma\right) \mid X_i = x\right)\\
&\quad - \lim_{x\nearrow 0} 2\mathbb{E}\left(Z_i^{(k)} \mid X_i = x\right)\mathbb{E}\left(Y_i - Z_i^\top\gamma \mid X_i = x\right)\Bigg)\\
&= \lim_{x\searrow 0}\mathbb{E}\left(2Z_i^{(k)}Z_i^{(l)} \mid X_i = x\right) - \lim_{x\searrow 0} 2\mathbb{E}\left(Z_i^{(k)} \mid X_i = x\right)\mathbb{E}\left(Z_i^{(l)} \mid X_i = x\right)\\
&\quad \lim_{x\nearrow 0}\mathbb{E}\left(2Z_i^{(k)}Z_i^{(l)} \mid X_i = x\right) - \lim_{x\nearrow 0} 2\mathbb{E}\left(Z_i^{(k)} \mid X_i = x\right)\mathbb{E}\left(Z_i^{(l)} \mid X_i = x\right)
\end{aligned}$$

hence

$$\begin{aligned}
&\mathrm{H}_{\gamma\mapsto\left(\lim_{x\searrow 0}\operatorname{Var}\left(Y_i - Z_i^\top\gamma \mid X_i = x\right) + \lim_{x\nearrow 0}\operatorname{Var}\left(Y_i - Z_i^\top\gamma \mid X_i = x\right)\right)}\\
&\equiv \lim_{x\searrow 0}\left(\operatorname{Cov}\left(Z_i^{(k)}, Z_i^{(l)} \mid X_i = x\right)\right)_{k,l\in\{1,...,p\}}\\
&\quad + \lim_{x\nearrow 0}\left(\operatorname{Cov}\left(Z_i^{(k)}, Z_i^{(l)} \mid X_i = x\right)\right)_{k,l\in\{1,...,p\}}.
\end{aligned} \tag{7.17}$$

We show that the first summand is positive semi-definite as this can be conducted analogously for the second summand. We have for $y \in \mathbb{R}^p \setminus \{0\}$

$$\begin{aligned}
&y^\top \lim_{x\searrow 0}\left(\operatorname{Cov}\left(Z_i^{(k)}, Z_i^{(l)} \mid X_i = x\right)\right)_{k,l\in\{1,...,p\}} y\\
&= \lim_{x\searrow 0} y^\top\Bigg(\mathbb{E}\Bigg(\left(Z_i^{(k)} - \mathbb{E}\left(Z_i^{(k)} \mid X_i = x\right)\right)\\
&\qquad\qquad \left(Z_i^{(l)} - \mathbb{E}\left(Z_i^{(l)} \mid X_i = x\right)\right) \mid X_i = x\Bigg)\Bigg)_{k,l\in\{1,...,p\}} y\\
&= \lim_{x\searrow 0} y^\top\mathbb{E}\left(\left(Z_i - \mathbb{E}(Z_i \mid X_i = x)\right)\left(Z_i - \mathbb{E}(Z_i \mid X_i = x)\right)^\top \mid X_i = x\right) y\\
&= \lim_{x\searrow 0}\mathbb{E}\left(\left(\left(Z_i - \mathbb{E}(Z_i \mid X_i = x)\right)^\top y\right)^2 \mid X_i = x\right)\\
&= \lim_{x\searrow 0}\mathbb{E}\left(\left(\sum_{k=1}^p y_k\left(Z_i^{(k)} - \mathbb{E}\left(Z_i^{(k)} \mid X_i = x\right)\right)\right)^2 \,\middle|\, X_i = x\right) \overset{(*)}{\geq} 0.
\end{aligned}$$

Hence, we know that the matrix $\lim_{x \searrow 0} \left(\mathrm{Cov}\left(Z_i^{(k)}, Z_i^{(l)} \mid X_i = x \right) \right)_{k,l \in \{1,\dots,p\}}$ is positive semi-definite, leading to the fact that we can apply the Cholesky decomposition to obtain

$$\lim_{x \searrow 0} \left(\mathrm{Cov}\left(Z_i^{(k)}, Z_i^{(l)} \mid X_i = x \right) \right)_{k,l \in \{1,\dots,p\}} = L^\top D L$$

for some orthogonal lower triangular matrix $L \in \mathbb{R}^{p \times p}$ and a diagonal matrix $D \in \mathbb{R}^{p \times p}$ having the real eigenvalues on its diagonal. Note that those eigenvalues are non-negative since the matrix is positive semi-definite and unequal to zero since we assumed the limit of the covariance matrix to be invertible. For this reason, D as well as $D^{\frac{1}{2}}$ are invertible, whereby $D^{\frac{1}{2}}$ is the diagonal matrix obtained by taking the square root of the diagonal entries of D. Therefore, $D^{\frac{1}{2}} L$ is invertible, which means that

$$y^\top L^\top D L y = y^\top L^\top D^{\frac{1}{2}} D^{\frac{1}{2}} L y = (D^{\frac{1}{2}} L y)^\top D^{\frac{1}{2}} L y = \left\| D^{\frac{1}{2}} L y \right\|_2^2 = 0 \quad \Leftrightarrow \quad y = 0.$$

Therefore, even a strict inequality holds in $(*)$, which proves that the matrix is positive definite. In an analogous way, we can prove that also the second summand in (7.17) is positive definite leading to the whole Hessian being positive definite. Finally, we can conclude that γ_n is a minimum.

□

8 Including Potentially Many Covariates

As already discussed in Section 3.5, the number of covariates has to be small in comparison to the number of observations as otherwise overfitting can cause a misleading estimator of the average treatment effect. However, having a large number of covariates that may even exceed the number of observations can occur quiet often in practice as someone may use large data sets or apply a large number of transformations on a low-dimensional setting. Therefore, studying the high-dimensional setting can be of great interest. For this reason, we want to consider the case that the number of covariates p is not fixed anymore and rather is allowed to grow as $n \to \infty$, i.e. the covariate dimension is p_n depending on n. By this point, the above studied theory is not appropriate anymore and collapses as it just produces an accurate estimator in low-dimensional settings. For this reason, the above theory needs to be enhanced. Indeed, [KR21] developed a theory that deals with a growing number of covariates by establishing an estimation in two steps: First, given the set of covariates, a small subset is selected which is strongly related to the outcome. In the second step, just the chosen covariates are included linearly into the local linear estimator of the average treatment effect. In fact, the main difference to the above theory is that we select a small subset of the covariates to be included in the regression, rather than just taking all of them into consideration. Under so-called approximate sparsity conditions (which informally means, that only a small number of covariates is relevant), [KR21] showed that this post-selection estimator is asymptotically normally distributed with bias and variance about the same as in the above case of a fixed covariate dimension. This chapter discusses the procedure of covariate selection and estimation as well as the results on the asymptotic behaviour, bias and variance. Yet, we do not provide proofs. Moreover, alternatives to the Lasso-based selection approach are developed and presented at the end of this chapter.

8.1 Covariate Selection

The easiest way to select a subset of the covariates is to use a fixed small subset $J = \{j_1, ..., j_s\} \subset \{1, ..., p_n\}$ such that $|J| \equiv s$ is very small in comparison to p_n. Then we can define the local linear estimator by linearly including these fixed chosen set of covariates as

$$\hat{\tau}_h(J) = e_2^\top \underset{(\theta,\gamma)\in\mathbb{R}^{4+s}}{\operatorname{argmin}} \sum_{i=1}^{n} K_h(X_i) \left(Y_i - V_i^\top \theta - Z_i(J)^\top \gamma\right)^2$$

whereby $Z_i(J) = (Z_i^{(1)}, ..., Z_i^{(j_s)})^\top$ is the vector of the selected covariates. Then, we could use the result of Chapter 7 as we have a fixed small number of covariates, leading to obtaining a consistent estimator under sufficient conditions. Nevertheless, using covariates that are strongly related to the outcome will deliver a more efficient estimation of the average treatment effect. As the covariates which are the most related to the outcome are not known immediately, they have to be determined in a data-driven way. One approach,

as described in [KR21], is to use a localized version of a Lasso regression in order to select the covariates. In particular, that means choosing a preliminary bandwidth b and a penalty parameter λ to solve the least squares problem

$$\left(\tilde{\theta}_n, \tilde{\gamma}_n\right) = \underset{(\theta,\gamma)\in\mathbb{R}^{4+p_n}}{\operatorname{argmin}} \sum_{i=1}^{n} K_b(X_i)\left(Y_i - V_i^\top\theta - (Z_i - \hat{\mu}_{Z,n})^\top\gamma\right)^2 + \lambda\sum_{k=1}^{p_n}\hat{w}_{n,k}|\gamma_k|$$

whereby

$$\hat{\mu}_{Z,n} = \frac{1}{n}\sum_{i=1}^{n} Z_i K_b(X_i) \quad \text{and} \quad \hat{w}_{n,k}^2 = \frac{b}{n}\sum_{i=1}^{n}\left(K_b(X_i)Z_i^{(k)} - \mu_{Z,n}^{(k)}\right)^2$$

are the local sample mean and variance of the covariates. The covariates that are the most related to the outcome are then indicated by the chosen subset

$$\hat{J}_n = \{k \in \{1,...,p_n\} \mid \tilde{\gamma}_n^{(k)} \neq 0\}.$$

The choices for b and λ have to be made provisionally since they do not occur in the asymptotic distribution of our estimator as we will see later.

8.2 Post-Lasso Estimator

After the previously described selection procedure of the covariates, the post-Lasso estimator of the average treatment effect is now defined by

$$\hat{\tau}_h(\hat{J}_n) = e_2^\top \underset{(\theta,\gamma)}{\operatorname{argmin}} \sum_{i=1}^{n} K_h(X_i)\left(Y_i - V_i^\top\theta - Z_i(\hat{J}_n)^\top\gamma\right)^2.$$

As already mentioned above, this estimator is asymptotically normally distributed. Yet, this is not true without further necessary conditions being satisfied. We do not want to state every condition formally and in detail. However, there is one key condition, which is of great importance and is the main difference to the low-dimensional case regarding the assumptions. In fact, this is the approximate sparsity condition.

Approximate sparsity condition Define the residual

$$r_i(J,h) = Y_i - V_i^\top\theta_0(J,h) - Z_i(J)^\top\gamma_0(J,h)$$

whereby

$$(\theta_0(J,h),\gamma_0(J,h)) = \underset{(\theta,\gamma)}{\operatorname{argmin}}\ \mathbb{E}\left(K_h(X_i)\left(Y_i - V_i^\top\theta - Z_i(J)^\top\gamma\right)^2\right).$$

We assume that $p_n \to \infty$ and

$$\max_{k=1,\dots,p_n} \left| \mathbb{E}\left(Z_{n,i}^{(k)} K_h(X_i) r_i(J_n, h) \right) \right| = O\left(\sqrt{\frac{\log p_n}{nh}} \right)$$

Also, suppose that this equation is true if we plug in the bandwidth b of the localized Lasso regression instead of h.

Intuitively, this assumption describes that only a small subset of the covariates is particularly relevant for the empirical analysis. The reason for this is that it states that every covariate which is not contained in the selection is locally uncorrelated to the regression error $r_i(J_n, h)$. Note that for every $k \in J_n$ the expression $\mathbb{E}\left(Z_{n,i}^{(k)} K_h(X_i) r_i(J_n, h) \right)$ vanishes anyways. Hence, as not chosen covariates are also not likely to have an impact on the regression error, it is reasonable to omit them in the empirical analysis.

Under this condition and additional regularity assumptions, [KR21] proved that

$$\frac{\sqrt{nh}\left(\hat{\tau}_h(\hat{J}_n) - \tau_Y - h^2 \mathcal{B}_n \right)}{\mathcal{S}_n} \xrightarrow{d} \mathcal{N}(0,1)$$

with bias and variance given by

$$\mathcal{B}_n = \frac{C_{\mathcal{B}}}{2}\left(\mu''_{\tilde{Y}+} - \mu''_{\tilde{Y}-} \right) + o(|J_n|^{\frac{1}{2}}) \quad \text{and} \quad \mathcal{S}_n^2 = \frac{C_{\mathcal{S}}}{f_X(0)}\left(\sigma^2_{\tilde{Y}+} + \sigma^2_{\tilde{Y}-} \right) + o(1)$$

whereby

$$\tilde{Y}_i = Y_i - Z(J_n)^\top \gamma_n \quad \text{with} \quad \gamma_n = \left(\sigma^2_{Z(J_n)-} + \sigma^2_{Z(J_n)+} \right)^{-1} \left(\sigma^2_{Z(J_n)Y-} + \sigma^2_{Z(J_n)Y+} \right)$$

and $C_{\mathcal{B}}$ as well as $C_{\mathcal{S}}$ are defined as in the previous chapter. The formulas of the bias and variance are indeed a natural generalization of the representations stated in the case of a fixed number of covariates due to $o(|J_n|^{\frac{1}{2}}) = o(1)$ for $p_n \equiv p$ and the definition of $\tilde{Y}_i$ being identical if Z_i is replaced by $Z_i(J_n)$.

8.3 Alternatives

As the Lasso procedure aims for selecting covariates that are strongly related to the outcome, it also seems to be reasonable to perform the selection according to the correlation of the covariates to the outcome. Indeed, we want to make suggestions for alternative covariate selection procedures in this subsection which will be examined within the next chapter by conducting simulations.

For every $k \in \{1, ..., p_n\}$ we can calculate the correlation of $Z_i^{(k)}$ and Y_i by

$$\operatorname{Corr}\left(Z_i^{(k)}, Y_i \right) = \frac{\operatorname{Cov}\left(Z_i^{(k)}, Y_i \right)}{\sigma_{Z_i^{(k)}} \sigma_{Y_i}} = \frac{\mathbb{E}\left(\left(Z_i^{(k)} - \mathbb{E}\left(Z_i^{(k)} \right) \right) \left(Y_i - \mathbb{E}(Y_i) \right) \right)}{\sqrt{\mathbb{E}\left(\left(Z_i^{(k)} - \mathbb{E}\left(Z_i^{(k)} \right) \right)^2 \right)} \sqrt{\mathbb{E}\left(\left(Y_i - \mathbb{E}(Y_i) \right)^2 \right)}},$$

whereby $\sigma_{Z_i^{(k)}}$ and σ_{Y_i} are the standard deviations of $Z_i^{(k)}$ and Y_i, respectively. Given a threshold ρ_k^b for each of the correlations of $Z_i^{(k)}$ and Y_i, we can now select all covariates that have a correlation with absolute value greater or equal ρ_k^b. That means the set J_n of selected indices is

$$J_n = \{k \in \{1, ..., p_n\} \mid |\rho_k| \geq \rho_k^b\}$$

whereby $\rho_k := \text{Corr}\left(Z_i^{(k)}, Y_i\right)$. Since it is infeasible to calculate ρ_k, we need to estimate it by the sample correlation, namely

$$\hat{\rho}_k := \frac{\frac{1}{n}\sum_{i=1}^n \left(Z_i^{(k)} - \overline{Z^{(k)}}\right)\left(Y_i - \overline{Y}\right)}{\hat{\sigma}_{Z_i^{(k)}}\hat{\sigma}_{Y_i}} = \frac{\sum_{i=1}^n \left(Z_i^{(k)} - \overline{Z^{(k)}}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^n \left(Z_i^{(k)} - \overline{Z^{(k)}}\right)^2 \sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2}}$$

with

$$\hat{\sigma}_{Z_i^{(k)}} = \sqrt{\frac{1}{n}\sum_{i=1}^n \left(Z_i^{(k)} - \overline{Z^{(k)}}\right)^2} \quad \text{and} \quad \hat{\sigma}_{Y_i} = \sqrt{\frac{1}{n}\sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2}$$

whereby $\overline{Z^{(k)}} = \frac{1}{n}\sum_{i=1}^n Z_i^{(k)}$ and $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ are the sample means. Then the data-driven version of J_n is

$$\hat{J}_n = \{k \in \{1, ..., p_n\} \mid |\hat{\rho}_k| \geq \rho_k^b\}.$$

Next, we want to discuss how ρ_k^b can be chosen.

Choice of correlation threshold Choosing ρ_k^b too low can lead to a large number of selected covariates, causing a misleading estimator, while choosing ρ_k^b too large can lead to missing out on beneficial covariates that may increase the accuracy of the estimation. Therefore, we have a trade-off situation and need to choose the threshold such that we take all relevant covariates into consideration and leave out those which just dilute our estimator's value. Let us assume that the covariates $Z_i^{(k)}$ are uncorrelated to Y_i. Then, we know that $\mathbb{E}(Z_i^{(k)}Y_i) = \mathbb{E}(Z_i^{(k)})\mathbb{E}(Y_i)$. Therefore, by applying Lindeberg-Lévy's Central Limit Theorem, we obtain

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n Z_i^{(k)}Y_i - \mathbb{E}\left(Z_i^{(k)}\right)\mathbb{E}(Y_i)\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma_k^2) \tag{8.1}$$

whereby σ_k^2 is defined by

$$\sigma_k^2 = \text{Var}\left(Z_i^{(k)}Y_i\right) = \mathbb{E}\left(\left(Z_i^{(k)}\right)^2 Y_i^2\right) - \mathbb{E}\left(Z_i^{(k)}\right)^2 \mathbb{E}(Y_i)^2.$$

If we now consider the correlation coefficient, write the covariance as

$$\text{Cov}\left(Z_i^{(k)}, Y_i\right) = \mathbb{E}\left(Z_i^{(k)}Y_i\right) - \mathbb{E}\left(Z_i^{(k)}\right)\mathbb{E}\left(Y_i\right),$$

and replace the term $\mathbb{E}(Z_i^{(k)}Y_i)$ by the sample mean, we obtain

$$\tilde{\rho}_k := \frac{\frac{1}{n}\sum_{i=1}^{n} Z_i^{(k)}Y_i - \mathbb{E}\left(Z_i^{(k)}\right)\mathbb{E}(Y_i)}{\sigma_{Z_i^{(k)}}\sigma_{Y_i}}.$$

Asymptotically, for large enough n, (8.1) states that we can find approximately 95 percent of all data points $\tilde{\rho}_k$ for an uncorrelated covariate $Z_i^{(k)}$ inside the interval

$$\left[-1.96\sigma_k/(\sqrt{n}\sigma_{Z_i^{(k)}}\sigma_{Y_i}), 1.96\sigma_k/(\sqrt{n}\sigma_{Z_i^{(k)}}\sigma_{Y_i})\right].$$

If we neglect the bias of the sample correlation coefficent, which is of order $O\left(\frac{1}{n}\right)$ [ZZW03], we can also approximately assume 95 percent of the data points $\hat{\rho}_k$ to be within that interval for large enough n, i.e.

$$\hat{\rho}_k \in \left[-1.96\sigma_k/(\sqrt{n}\sigma_{Z_i^{(k)}}\sigma_{Y_i}), 1.96\sigma_k/(\sqrt{n}\sigma_{Z_i^{(k)}}\sigma_{Y_i})\right].$$

Since we want to confirm this approach in a data-driven way by conducting a simulation, it is appropriate to not be too strict about the bias here. Now we can formulate this as a statistical test with null hypothesis of $Z_i^{(k)}$ being uncorrelated to Y_i. We want to test the null hypothesis at a significance level of $\alpha = 0.05$. For this reason, we reject the null hypothesis if

$$|\hat{\rho}_k| \geq 1.96\sigma_k/(\sqrt{n}\sigma_{Z_i^{(k)}}\sigma_{Y_i}).$$

Hence, we can argue that choosing the correlation threshold $\rho_k^b = 1.96\sigma_k/(\sqrt{n}\sigma_{Z_i^{(k)}}\sigma_{Y_i})$ seems to be reasonable as, then, almost all of the uncorrelated covariates will be dropped by the selection algorithm. Yet, this just guarantees that unfavorable covariates are likely to be excluded, but does not ensure that correlated covariates are actually selected. But, this is only partially a problem since we want to be very restrictive when it comes to selecting covariates in order to keep the number of chosen covariates low. In fact, the priority is to reject unhelpful covariates rather than taking beneficial covariates into consideration.

Now, given a data set, it is still infeasible to determine ρ_k^b, which is why we need the sample threshold

$$\hat{\rho}_k^b = \frac{1.96\hat{\sigma}_k}{\sqrt{n}\hat{\sigma}_{Z_i^{(k)}}\hat{\sigma}_{Y_i}}$$

whereby $\hat{\sigma}_{Z_i^{(k)}}$ as well as $\hat{\sigma}_{Y_i}$ are defined as above and

$$\hat{\sigma}_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Z_i^{(k)}\right)^2 Y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} Z_i^{(k)}\right)^2\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)^2}.$$

Bonferroni correction However, when we have the situation of p different covariates, this hypothesis test is conducted p times for each of those covariates within one simulation run. This, indeed, causes an accumulation of type 1 errors and as a consequence, we

do not reach an overall significance level of 0.05. For this reason, we need to adjust our above correlation threshold in order to compensate that. Therefore, we now want the null hypothesis to be that all of the p covariates $Z_i^{(k)}$ are uncorrelated to the outcome. Then, the null hypothesis is rejected when one of the individual null hypothesis of $Z_i^{(k)}$ being uncorrelated to the outcome is rejected. This situation is indeed suitable for performing a Bonferroni correction in order to compensate the increasing likelihood of incorrectly rejecting a null hypothesis, which means falsely selecting a covariate. As our goal is to obtain an overall significance level of $\alpha = 0.05$, the Bonferroni correction suggests to test each individual null hypothesis (of $Z_i^{(k)}$ being uncorrelated to the outcome) at a significance level of

$$\alpha_{\text{ind}}^{p} = \frac{0.05}{p}.$$

Then we choose ω_p such that the value $\hat{\rho}_k$ for an uncorrelated $Z_i^{(k)}$ falls inside the interval

$$\hat{\rho}_k \in \left[-\omega_p \sigma_k / (\sqrt{n}\sigma_{Z_i^{(k)}}\sigma_{Y_i}), \omega_p \sigma_k / (\sqrt{n}\sigma_{Z_i^{(k)}}\sigma_{Y_i})\right],$$

with percentage

$$(1 - \alpha_{\text{ind}}^{p}) \cdot 100.$$

Then, we define our Bonferroni corrected correlation threshold as

$$\hat{\rho}_k^{b,p,\text{corrected}} := \frac{\omega_p \hat{\sigma}_k}{\sqrt{n}\hat{\sigma}_{Z_i^{(k)}}\hat{\sigma}_{Y_i}}.$$

Later on, our simulations will incorporate a number of 200 or 1593 covariates, respectively, depending on the setting. Therefore, applied on these specific cases we obtain the following: In the case of $p = 200$, we get a corrected correlation threshold of approximately

$$\hat{\rho}_k^{b,200,\text{corrected}} = \frac{3.66\hat{\sigma}_k}{\sqrt{n}\hat{\sigma}_{Z_i^{(k)}}\hat{\sigma}_{Y_i}},$$

while we obtain in the case of $p = 1593$ a value of about

$$\hat{\rho}_k^{b,1593,\text{corrected}} = \frac{4.16\hat{\sigma}_k}{\sqrt{n}\hat{\sigma}_{Z_i^{(k)}}\hat{\sigma}_{Y_i}},$$

in order to reach an overall significance level of 0.05.

Next, we describe all selection procedures that we want to examine in the remainder of this thesis.

8.3.1 Selection by Correlation Threshold

We choose all covariates $Z_i^{(k)}$ that satisfy the condition $|\hat{\rho}_k| \geq \hat{\rho}_k^{b,p,\text{corrected}}$. This ensures that covariates uncorrelated to the outcome are very likely to be rejected. One potential weakness might be that the threshold contains the factor $\frac{1}{\sqrt{n}}$ and, therefore, drops with

increasing n. For this reason, especially for large data sets, the number of selected covariates may become too large. In the remainder of this thesis, we will refer to this selection procedure by (CCT), which stands for "Calculated Correlation Threshold".

8.3.2 Selection by Correlation Threshold with Simple Deletion

With this procedure we want to address the concern that the number of chosen covariates may become too large for growing n. As in the previous procedure, we first want to select all covariates $Z_i^{(k)}$ that satisfy the condition $|\hat{\rho}_k| \geq \hat{\rho}_k^{b,p,\text{corrected}}$. Now, let $Z(\hat{J}) = \{Z_i^{(j_1)}, ..., Z_i^{(j_s)}\}$ be the set of these chosen covariates with $j_1 < j_2 < ... < j_s$. Indeed, this set now might contain several covariates that are pairwise correlated. In the case of a correlated pair of two selected covariates, we do not want to keep both as this might significantly increase the number of chosen covariates while only causing a marginal increase in new information being provided. In fact, let $k \neq l$, then we calculate the sample correlation coefficient of $Z_i^{(j_k)}$ and $Z_i^{(j_l)}$

$$\hat{\eta}_{kl} := \frac{\sum_{i=1}^n \left(Z_i^{(j_k)} - \overline{Z^{(j_k)}}\right)\left(Z_i^{(j_l)} - \overline{Z^{(j_l)}}\right)}{\sqrt{\sum_{i=1}^n \left(Z_i^{(j_k)} - \overline{Z^{(j_k)}}\right)^2 \sum_{i=1}^n \left(Z_i^{(j_l)} - \overline{Z^{(j_l)}}\right)^2}}$$

with $\overline{Z^{(j_k)}}$ and $\overline{Z^{(j_l)}}$ being the sample means as above. Given $k < l$ and a threshold $\hat{\eta}_{kl}^{b,p} > 0$, we want to drop one of those two covariates if $|\hat{\eta}_{kl}| \geq \hat{\eta}_{kl}^{b,p}$. Indeed, in that case, we drop the covariate with the smaller correlation to the outcome, i.e. if $|\hat{\rho}_{j_k}| < |\hat{\rho}_{j_l}|$, we delete $Z_i^{(j_k)}$ from the set of selected covariates, otherwise we delete $Z_i^{(j_l)}$.

However, we still need to determine the threshold $\hat{\eta}_{kl}^{b,p}$. It needs to be chosen such that all uncorrelated pairs are very likely to lie below that threshold. Therefore, we can derive the threshold with the same reasoning as for the above correlation threshold $\hat{\rho}_k^{b,p,\text{corrected}}$ by using Lindeberg-Lévy's Central Limit Theorem. This leads to

$$\hat{\eta}_{kl}^{b,p} := \frac{\xi_p \hat{\sigma}_{kl}}{\sqrt{n}\hat{\sigma}_{Z_i^{(j_k)}}\hat{\sigma}_{Z_i^{(j_l)}}}$$

whereby $\hat{\sigma}_{Z_i^{(j_k)}}$ as well as $\hat{\sigma}_{Z_i^{(j_l)}}$ are the sample standard deviations defined as above,

$$\hat{\sigma}_{kl} = \sqrt{\frac{1}{n}\sum_{i=1}^n \left(Z_i^{(j_k)}\right)^2 \left(Z_i^{(j_l)}\right)^2 - \left(\frac{1}{n}\sum_{i=1}^n Z_i^{(j_k)}\right)^2 \left(\frac{1}{n}\sum_{i=1}^n Z_i^{(j_l)}\right)^2}$$

and ξ_p the factor chosen such that the covariate deletion reaches an overall significance level of 0.05. This significance level, in fact, can be reached by choosing ξ_p with a similar reasoning as ω_p: Let s be the number of chosen covariates by (CCT). Then, we consider the null hypothesis to be that every of the $\frac{s(s-1)}{2}$ pairs of covariates examined by the deletion algorithm is uncorrelated. It is rejected, if any of the individual null hypothesis

of $Z_i^{(j_k)}$ and $Z_i^{(j_l)}$ being uncorrelated is rejected. The Bonferroni correction now suggest to set the significance level of each of those individual tests to

$$\alpha_{\text{ind}}^p = \frac{0.05}{0.5 \cdot s(s-1)} = \frac{0.1}{s(s-1)}.$$

As a consequence, we choose ξ_p such that

$$(1 - \alpha_{\text{ind}}^p) \cdot 100$$

percent of the values $\hat{\eta}_{k,l}$ for uncorrelated $Z_i^{(j_k)}$ and $Z_i^{(j_l)}$ lie in the interval

$$\left[\frac{-\xi_p \sigma_{kl}}{\sqrt{n}\sigma_{Z_i^{(j_k)}}\sigma_{Z_i^{(j_l)}}}, \frac{\xi_p \sigma_{kl}}{\sqrt{n}\sigma_{Z_i^{(j_k)}}\sigma_{Z_i^{(j_l)}}} \right],$$

whereby $\sigma_{kl} = \sqrt{\text{Var}(Z_i^{(j_k)} Z_i^{(j_l)})}$ and $\sigma_{Z_i^{(j_k)}}$ as well as $\sigma_{Z_i^{(j_l)}}$ are the standard deviations of $Z_i^{(j_k)}$ and $Z_i^{(j_l)}$, respectively. Again, this consideration neglects the error of the sample correlation coefficient.

Overall, this selection procedure might resolve the potential issue of too many selected covariates, especially, since the "deletion threshold" $\hat{\eta}_{kl}^{b,p}$ also drops with growing n. Therefore, it seems to compensate the effect that more covariates are chosen by (CCT) with increasing n. In the subsequent content, we will refer to this selection procedure by (CCTSD), which stands for "Calculated Correlation Threshold with Simple Deletion".

8.3.3 Selection by Correlation Threshold with Advanced Deletion

One issue of the previous deletion strategy might be that it deletes too many covariates and, therefore, wastes potentially helpful information. Why this is the case, can be seen in the directed graph $G = (V, E)$ illustrated in Figure 10. We use the terminology of the previous selection procedure. Then, V is the set of indices of the selected covariates chosen by (CCT), i.e.

$$V = \hat{J} = \{j_1, j_2, ..., j_s\}.$$

We define the set of edges $E \subseteq V \times V$ as following:

$$\begin{aligned} &(j_l, j_k) \in E \\ :\Leftrightarrow\ &\Big(j_k \neq j_l \ \wedge\ |\hat{\eta}_{kl}| \geq \hat{\eta}_{kl}^{b,p} \ \wedge\ \big(|\hat{\rho}_{j_k}| < |\hat{\rho}_{j_l}| \ \vee\ (|\hat{\rho}_{j_l}| = |\hat{\rho}_{j_k}| \ \wedge\ j_k > j_l) \big) \Big). \end{aligned}$$

Therefore, an edge $(j_l, j_k) \in E$ means that $Z_i^{(j_k)}$ and $Z_i^{(j_l)}$ are a pair of covariates with $Z_i^{(j_k)}$ being a candidate for deletion by the deletion approach described in Section 8.3.2. All nodes which are marked red in the graph indeed would get removed by the previous deletion algorithm, while just the green marked nodes would remain in the set of selected covariates. Our goal was to dissolve pairs of covariates which are correlated, whereby the

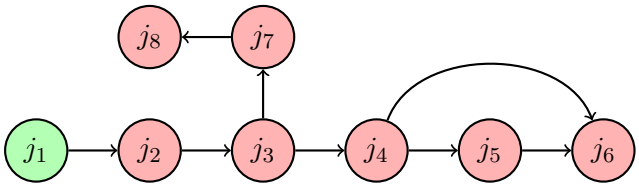


Figure 10: Graph for simple covariate deletion

covariate with the larger correlation to the outcome is kept, while the other one is deleted. However, this can be accomplished by deleting less covariates. In fact, we want to take into consideration when a covariate has already been deleted and, then, not delete covariates which would get dropped because of being correlated to that already deleted covariate. This seems to be a reasonable approach since correlation is not transitive, which indeed means that if A is correlated to B and B is correlated to C, it does not necessarily have to be the case that A is correlated to C. Hence, as a deletion of B would already dissolve the correlation dependencies when A is not correlated to C, there is no need to additionally delete either A or C. Therefore, we want to establish an alternative deletion procedure to the above in the following. Therefore, let $Z(\hat{J}) = \{Z_i^{(j_1)}, ..., Z_i^{(j_s)}\}$ with $j_1 < j_2 < ... < j_s$ be the set of those chosen covariates by the selection approach (CCT) described in Section 8.3.1. Next, we sort the covariates according to the absolute value of their correlation to the outcome. So, let

$$Z_i^{\text{sort}} = \left(Z_i^{(j_{u_1})}, Z_i^{(j_{u_2})}, ..., Z_i^{(j_{u_s})}\right)$$

be the tuple of sorted covariates, i.e. $|\hat{\rho}_{j_{u_1}}| \geq |\hat{\rho}_{j_{u_2}}| \geq ... \geq |\hat{\rho}_{j_{u_s}}|$. In case that $|\hat{\rho}_{j_k}| = |\hat{\rho}_{j_l}|$ for some $k < l$, we sort according to the indices and list $Z_i^{(j_k)}$ before $Z_i^{(j_l)}$ in the sorted tuple. Now we delete covariates according to the order of this sorted tuple. In fact, let

$$D := \{(j_k, j_l) \mid |\hat{\eta}_{kl}| \geq \hat{\eta}_{kl}^{b,p}, k \neq l\}$$

be the set of index pairs, whose according covariate pairs exceed the correlation "deletion threshold". In addition, define

$$D_v := D \cap \{(j_{u_v}, j_l) \mid l \in \{1, ..., s\}\}$$

for $v = 1, ..., s$. The deletion procedure now processes D_v in step v. That means, it begins with D_1, then D_2, and so on, until it ends with D_s. In fact, in step v, we delete all covariates $Z_i^{(d)}$ with its index d contained in

$$\tilde{D}_v := \begin{cases} \varnothing & \text{if } j_{u_v} \in \bigcup_{w=1}^{v-1} \tilde{D}_w \\ \{d \mid (j_{u_v}, d) \in D_v\} & \text{otherwise.} \end{cases}$$

Indeed, we just want to delete covariates if $Z_i^{(j_{u_v})}$ has not already been deleted yet in one of the previous steps $1, ..., v-1$. This is checked by the condition $j_{u_v} \notin \bigcup_{w=1}^{v-1} \tilde{D}_w$, which

means that j_{u_v} is not contained in the sets of indices of deleted covariates from previous steps. If j_{u_v}, in turn, is contained in one of these sets, no covariates are delete in step v.

Since the description of our deletion algorithm is very technical, we want to illustrate the result of it applied on the above graph under the assumption that $Z_i^{\text{sort}} = (Z_i^{(j_1)}, Z_i^{(j_2)}, ..., Z_i^{(j_8)})$. This is visualized in Figure 11. As we can already see, this deletion

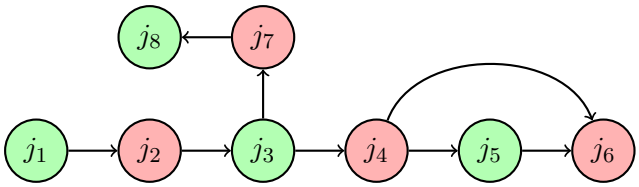


Figure 11: Graph for advanced covariate deletion

approach tends to delete less covariates. It is important to highlight that this algorithm keeps the covariates with the highest absolute value of correlation to the outcome under the condition of dissolving all pairs of correlated covariates. In the remainder of this thesis, we will refer to this selection procedure by (CCTAD), which stands for "Calculated Correlation Threshold with Advanced Deletion".

The next chapter covers test results of simulations examining all three of our alternative covariate selection methods.

9 Simulations

In this section we want to conduct a simulation in order to test the selection procedures described as alternatives to the Lasso-based selection in the previous chapter. We want to do that in two scenarios: On the one hand, in a setting with generated data and, on the other hand, in a setting with a real data set about the Austrian labor market.

9.1 Implementation

The simulations are coded using the language `R` and implement the covariate selection as well as the post-selection estimation with linearly included covariates as described throughout this thesis. In order to realize this, a bandwidth h needs to be chosen. This can be done by methods either stated in [CCFT19] or in [AK18], which are implemented in the R packages `rdrobust` [CCFT22] or `RDHonest` [Kol23], respectively. We want to conduct the first simulation using both frameworks in order to examine if the bandwidth choice implementation causes a huge difference in the result. However, then, the following simulations only use `RDHonest`.

Numerical invertibility The simulation with `RDHonest` requires to do the covariate adjustment manually since there is no already implemented method for that. `rdrobust` already contains implemented functionalities for including covariates. Nevertheless, the covariate adjustment is implemented accordingly to the procedure described in Definition 7.2 in the case of both frameworks. Particularly, this includes computing the inverse of $\sigma_{Z+} + \sigma_{Z-}$. This can be numerically impossible if the number of chosen covariates is large and if there are pairs of covariates with a relatively high correlation. If this case occurs, we solve it by deleting d covariates for a number $d \in \mathbb{N}$ such that we dissolve pairs of covariates with a high correlation. The settings in which this error of not being numerically invertible occurred, will be marked below. Then, we respectively choose the number d such that the numerical invertibility issue was resolved. However, it is important to mention that we only had to apply this fix in settings constructed for comparison purposes with a large number of chosen covariates. Therefore, it does not dilute the results of our selection procedures of interest.

9.2 Generated Data

The advantage of generating data is that, as we exactly know the average treatment effect and the influence of the covariates on the outcome, we can assess the accuracy of our estimation. We want to run the simulation on the same data model as in [KR21]. In fact, we have a RDD setting with $p = 200$ covariates and an average treatment effect of

$\tau_Y = 0.02$, which is described as follows:

$$X \sim 2 \cdot \text{beta}(2,4) - 1, \quad T = \mathbb{1}(X \geq 0), \quad (\varepsilon, Z^\top)^\top \sim \mathcal{N}(0, \Sigma), \quad \Sigma = \begin{pmatrix} \sigma_\varepsilon^2 & v^\top \\ v & \sigma_Z^2 I_{200} \end{pmatrix},$$

$$Y = \varepsilon + \begin{cases} 0.36 + 0.96X + 5.47X^2 + 15.28X^3 + 15.87X^4 + 5.14X^5 + 0.22Z^\top\alpha, & \text{if } T = 0 \\ 0.38 + 0.62X - 2.84X^2 + 8.42X^3 - 10.24X^4 + 4.31X^5 + 0.28Z^\top\alpha, & \text{if } T = 1 \end{cases}$$

whereby $I_{200} \in \mathbb{R}^{200\times 200}$ denotes the identity matrix, $\sigma_\varepsilon^2 = 0.1295^2, \sigma_Z^2 = 0.1353^2$, $v \in \mathbb{R}^{200}$ a vector whose k-th entry equals $v_k = 0.8\sqrt{6}\sigma_\varepsilon^2(\pi k)^{-1}$ and $\alpha \in \mathbb{R}^{200}$ is a vector such that the k-th component equals $\alpha_k = 2k^{-2}$. The definition of α suggests that the impact of the covariates falls as k decreases. Therefore, the covariates $Z = (Z^{(1)}, ..., Z^{(200)})$ are ordered according to their relevance with $Z^{(1)}$ being the most important one, while $Z^{(200)}$ being almost negligible. We run the simulation for two different choices of the sample size n. First, we run the simulation for a sample size of $n = 1000$ with 10.000 Monte Carlo replications. Secondly, we set the sample size to $n = 10,000$ whereby we also run 10,000 Monte Carlo replications. In the simulation, we perform the estimation for different choices of covariates, namely:

- Setting 1 (CCT): Selection of all covariates $Z^{(k)}$ with $|\hat{\rho}_k| \geq \hat{\rho}_k^{b,200,\text{corrected}}$ (as described in Section 8.3.1)
- Setting 2 (CCTSD): Selection of all covariates $Z^{(k)}$ with $|\hat{\rho}_k| \geq \hat{\rho}_k^{b,200,\text{corrected}}$ and then applying the simple deletion procedure (as described in Section 8.3.2)
- Setting 3 (CCTAD): Selection of all covariates $Z^{(k)}$ with $|\hat{\rho}_k| \geq \hat{\rho}_k^{b,200,\text{corrected}}$ and then applying the advanced deletion procedure (as described in Section 8.3.3)
- Setting 4: Selection of all covariates $Z^{(k)}$ with $|\hat{\rho}_k| \geq 0.2$
- Setting 5: Selection of all covariates $Z^{(k)}$ with $|\hat{\rho}_k| \geq 0.1$
- Setting 6: Selection of all covariates $Z^{(k)}$ with $|\hat{\rho}_k| \geq 0.05$ (This setting, in the case of $n = 1,000$, required to delete 15 covariates in order to obtain numerical invertibility. This was not necessary for $n = 10,000$.)
- Setting 7: Selection of all covariates $Z^{(k)}$ with $|\hat{\rho}_k| \geq 0.02$ (This setting, in the case of $n = 1,000$, required to delete 45 covariates in order to obtain numerical invertibility. This was not necessary for $n = 10,000$.)

The Settings 4 to 7 allow us to assess the appropriateness of our derived calculated correlation threshold and whether it actually delivers a favorable lower bound. Additionally, we also want to simulate the case when we take a fixed and predetermined subset of covariates:

- Setting 8: Selection of no covariates

- Setting 9: Selection of most important covariate
- Setting 10: Selection of 10 most important covariates
- Setting 11: Selection of 30 most important covariates
- Setting 12: Selection of 50 most important covariates
- Setting 13: Selection of the "optimal" covariate $Z^\top \alpha$

Of course, the selections in the Settings 8 to 13 are infeasible in practice since the most important covariates are unknown. However, we want to use these settings with a fixed choice of covariates for the sake of comparison in order to have benchmark values. In particular, Setting 13 provides us with desirable values for the bias, standard error, standard deviation and the confidence interval. The best case, therefore, would be to approximately obtain these optimal values for Setting 1, 2 and 3 in order to conclude that our selection procedures work appropriately.

For each of those settings, the following values will be determined:

- Number of covariates: For every of the Monte Carlo replications, the simulation stores the number of selected covariates by each of the procedures. Then, the average over all executions is calculated, respectively.
- Bias: This is the difference of the sample mean of the estimator and the true parameter 0.02. The sample mean is determined by averaging the estimated average treatment effect over all executions.
- Standard deviation (SD): It is calculated by averaging the squared difference of the estimated values of the average treatment effect and the sample mean of the estimator.
- Average standard error (SE): This is the value calculated internally by the respective `R` package (`rdrobust`, `RDHonest`) averaged over all replications.
- CI Length: This is the length of the confidence interval provided by the respective `R` package (`rdrobust`, `RDHonest`) averaged over all replications.
- Coverage: For every execution it is determined whether the true parameter 0.02 lies within the confidence interval. Then, considered among all executions, the coverage is the probability that the true parameter was covered by the confidence interval.

9.2.1 Moderate Sample Size

First, we want to run the simulation on the above data model with a sample size of $n = 1,000$ with 10,000 Monte Carlo replications. We provide the results for both frameworks `RDHonest` and `rdrobust`.

Simulation results for `RDHonest` The results of the simulation are listed in Table 1. First, we want to have a look at the results for the selection procedures taking a

Covariate Selection	Cov	Bias	SD	Avg. SE	CI Length	Coverage
(CCT)	2.20	0.0068	0.0395	0.0378	0.1665	95.4
(CCTSD)	2.15	0.0069	0.0398	0.0380	0.1675	95.5
(CCTAD)	2.15	0.0069	0.0398	0.0380	0.1675	95.5
Corr. threshold of 0.2	1.98	0.0069	0.0400	0.0384	0.1691	95.4
Corr. threshold of 0.1	4.50	0.0064	0.0376	0.0349	0.1547	94.9
Corr. threshold of 0.05	16.04	0.0068	0.0379	0.0312	0.1416	92.7
Corr. threshold of 0.02	66.61	0.0090	0.0446	0.0181	0.1017	74.1
Fixed: no cov.	-	0.0094	0.0667	0.0651	0.2854	96.2
Fixed: most imp. cov.	-	0.0076	0.0445	0.0430	0.1890	95.5
Fixed: 10 most imp. cov.	-	0.0066	0.0360	0.0317	0.1422	94.1
Fixed: 30 most imp. cov.	-	0.0077	0.0374	0.0261	0.1232	88.4
Fixed: 50 most imp. cov.	-	0.0088	0.0401	0.0213	0.1089	81.0
Fixed: optimal cov.	-	0.0067	0.0375	0.0362	0.1594	95.3

Table 1: Simulation results using `RDHonest` with a sample size of 1,000 based on 10,000 Monte Carlo replications

fixed threshold for the correlation. The bias and standard deviation do slightly increase as the threshold decreases from 0.1 to 0.02, that is with growing number of covariates that have been chosen. Even more important, we can see that the average standard error of the estimator decreases the lower the threshold gets. In contrast to that, the standard deviation does not decrease, rather tends to even increase, leading to a growing discrepancy between the average standard error and the standard deviation. This seems to suggest that the framework's standard error underestimates the true standard deviation with increasing number of covariates which indeed is due to overfitting. In the extreme case of selecting covariates with correlation greater than a threshold of 0.02, we can see that over 66 covariates have been taken into consideration on average, leading to a highly underestimated standard error. This, in turn, implies a too small confidence interval. For this reason, we can see how the coverage plummets with decreasing threshold. However, the results for a fixed correlation threshold of 0.2 and 0.1 are both desirable as the bias and standard deviation are lower as in the case of no covariates while only having a marginal drop in coverage of approximately one percent.

The phenomenon of a progressively strong downward bias in the standard error can also be seen in the case of a fixed choice of covariates. With increasing number of chosen covariates, we can see the difference between the standard deviation and average standard error getting bigger, leading to increasingly inaccurate confidence intervals. Therefore, our simulation confirmed that the approach described throughout this thesis just works well in low-dimensional settings. Nonetheless, the results also seem to confirm that taking a small and well selected number of covariates into consideration can improve the bias and standard deviation of the estimator while only having a slight decline in the CI coverage. Indeed, comparing the results of the choice via the calculated correlation threshold (CCT)

to the case without covariates shows that the coverage just falls by 0.8 percent while significantly improving the standard deviation. Also, the values of the bias and standard deviation are even close to those obtained in the case of choosing the obstacle optimal linear combination of covariates.

Moreover, the simulation results show that (CCT) seems to select a rather small number of covariates. Also the selected covariates tend to be not pairwise correlated as the simple and advanced deletion procedure of (CCTSD) and (CCTSD) just very rarely delete some of the chosen covariates. We can see that the average number of chosen covariates decreased from 2.20 to just 2.15. For this reason, the results regarding bias, standard deviation, average standard error, CI length and coverage are nearly the same. To see which of the covariates have been rarely deleted, we refer to Table 2. It shows the percentages with which a covariate was chosen by the respective selection procedures.

Covariate	(CCT)	(CCTAD)	(CCTSD)
Number 1	100.00	99.49	99.48
Number 2	98.86	95.04	95.04
Number 3	19.38	18.28	18.28
Number 4	1.53	1.28	1.28
Number 5	0.14	0.10	0.10
Number 6	0.03	0.00	0.00
Number 7	0.01	0.00	0.00
Number 10	0.01	0.00	0.00
All others	0.00	0.00	0.00

Table 2: Percentages of the covariates being chosen by the different selection procedures with a sample size of 1,000 based on 10,000 Monte Carlo replications

Overall, we can conclude that all three selection approaches (CCT), (CCTSD) and (CCTAD) seem to deliver accurate estimations with desirable inference results that can even keep up with the optimal benchmark results.

Simulation results for `rdrobust` Table 3 shows the simulation results. The values in the table are different from the results using `RDHonest`, which is reasonable due to the fact that other methods for bandwidth selection and calculation of the standard error are used. However, we can see that the behaviour of the values is very similar. The only major difference is that the standard error of `rdrobust` seems to underestimate the true standard deviation even more leading to overall shorter confidence intervals and lower CI coverages. Similarly to the behaviour of `RDHonest`, the downward bias in the standard error even grows with the number of selected covariates, which is the reason for rapidly declining confidence intervals. In particular, we can see that a fixed correlation threshold of 0.02 leads to a coverage of just 53.5 percent. Nonetheless, concerning our selection procedures (CCT), (CCTSD) and (CCTAD), we can conclude the same consequences as above. Indeed, just the drop in the CI coverage compared to the case of no covariates is larger which seems to be owing to the covariates scaling up the effect of `rdrobust`

Covariate Selection	Cov	Bias	SD	Avg. SE	CI Length	Coverage
(CCT)	2.21	0.0150	0.0331	0.0285	0.1117	87.0
(CCTSD)	2.16	0.0150	0.0333	0.0287	0.1125	87.0
(CCTAD)	2.16	0.0150	0.0333	0.0287	0.1125	87.0
Corr. threshold of 0.2	1.98	0.0150	0.0336	0.0290	0.1136	87.1
Corr. threshold of 0.1	4.52	0.0147	0.0315	0.0265	0.1037	85.7
Corr. threshold of 0.05	16.09	0.0136	0.0323	0.0249	0.0977	82.8
Corr. threshold of 0.02	66.65	0.0095	0.0535	0.0165	0.0648	53.5
Fixed: no cov.	-	0.0168	0.0573	0.0524	0.2053	91.8
Fixed: most imp. cov.	-	0.0153	0.0371	0.0328	0.1286	88.7
Fixed: 10 most imp. cov.	-	0.0144	0.0304	0.0245	0.0960	83.4
Fixed: 30 most imp. cov.	-	0.0129	0.0326	0.0220	0.0863	77.1
Fixed: 50 most imp. cov.	-	0.0112	0.0375	0.0193	0.0757	66.8
Fixed: optimal cov.	-	0.0151	0.0312	0.0269	0.1056	86.0

Table 3: Simulation results using `rdrobust` with a sample size of 1,000 based on 10,000 Monte Carlo replications

underestimating the true standard deviation even more.

9.2.2 Large Sample Size

As already suspected in Section 8.3.1, the number of covariates chosen by (CCT) might increase with growing n. This is suggested by the fact that the calculated correlation threshold includes the factor $n^{-\frac{1}{2}}$. For this reason, we also want to examine our selection procedures in a setting with a sample size of $n = 10,000$ and 10,000 Monte Carlo Replications. Since `RDHonest` delivered the more favorable results with more accurate confidence intervals above, we just want to stick to this package now and neglect the simulation with `rdrobust`. The simulation results can be seen in Table 4.

Covariate Selection	Cov	Bias	SD	Avg. SE	CI Length	Coverage
(CCT)	6.89	0.0030	0.0139	0.0137	0.0601	96.4
(CCTSD)	6.84	0.0030	0.0139	0.0137	0.0602	96.5
(CCTAD)	6.84	0.0030	0.0139	0.0137	0.0602	96.5
Corr. threshold of 0.2	2.00	0.0033	0.0154	0.0153	0.0669	96.5
Corr. threshold of 0.1	3.72	0.0031	0.0144	0.0143	0.0627	96.6
Corr. threshold of 0.05	7.54	0.0030	0.0139	0.0137	0.0598	96.3
Corr. threshold of 0.02	31.10	0.0030	0.0135	0.0129	0.0569	95.9
Fixed: no cov.	-	0.0051	0.0251	0.0249	0.1089	96.5
Fixed: most imp. cov.	-	0.0037	0.0171	0.0170	0.0744	96.7
Fixed: 10 most imp. cov.	-	0.0030	0.0137	0.0135	0.0590	96.3
Fixed: 30 most imp. cov.	-	0.0030	0.0135	0.0129	0.0566	95.9
Fixed: 50 most imp. cov.	-	0.0031	0.0135	0.0125	0.0554	95.2
Fixed: optimal cov.	-	0.0031	0.0145	0.0144	0.0631	96.5

Table 4: Simulation results using `RDHonest` with a sample size of 10,000 based on 10,000 Monte Carlo replications

First of all, we can see that the phenomenon of a progressively strong downward bias in the average standard error with growing number of covariates is not visible as clearly as before because the numbers of selected covariates are overall smaller in relation to the sample size. The results suggest that even around 30 selected covariates do not lead to an extensively inaccurate confidence interval and an unfavorable coverage. Yet, we can still see that the discrepancy between the average standard error and the true standard deviation slightly increases with growing number of included covariates leading to a marginally visible progressively strong downward bias. Moreover, we can see that including more covariates than our three selection procedures (CCT), (CCTSD) and (CCTAD) does not improve the bias and standard deviation significantly, which suggests that it is unnecessary to include those further covariates.

Secondly, the results seem to confirm that (CCT) chooses more covariates with growing n. However, 6.89 selected covariates on average is still a reasonable number compared to the sample size of $10,000$. This is also reflected by the fact that the coverage is only 0.1 percent lower than in the case of including no covariates. Again, all selected covariates seem to be pairwise uncorrelated since the number of selected covariates by (CCTSD) and (CCTAD) is in average just 0.05 lower.

Overall, (CCT), (CCTSD) and (CCTAD) again deliver desirable estimation results.

9.3 Generated Data with Duplicated Covariates

In the simulations conducted above, we can see that (CCTSD) and (CCTAD) delete a rather low number of covariates. Now, we want to conduct another simulation which demonstrates more clearly that our simple and advanced deletion procedure work appropriately. Therefore, we construct a data model similar to the one above, with the only difference that we just take the 20 most important covariates and repeat them for 10 times. This results in a set of 200 covariates with just 20 different covariates, whereby each of them occurs 10 times. Indeed, this scenario might occur in real data sets, e.g. when the data is not cleansed prior to the simulation or if correlated covariates are incorporated.

The data model is defined exactly as above with the only adjustment that we now take

$$Z_{\text{duplicates}} = (\underbrace{Z^{(1)}, ..., Z^{(20)}}_{\text{Number 1}}, \underbrace{Z^{(1)}, ..., Z^{(20)}}_{\text{Number 2}}, \ \ldots \ , \underbrace{Z^{(1)}, ..., Z^{(20)}}_{\text{Number 10}})$$

as set of covariates (instead of Z). We conduct our simulation using `RDHonest` with a sample size of $n = 1,000$ and $10,000$ Monte Carlo replications. We will provide estimation results for the following settings:

- Setting 1 (CCT): Selection of all covariates $Z^{(k)}_{\text{duplicates}}$ with $|\hat{\rho}_k| \geq \hat{\rho}_k^{b,200,\text{corrected}}$ (as described in Section 8.3.1)
- Setting 2 (CCTSD): Selection of all covariates $Z^{(k)}_{\text{duplicates}}$ with $|\hat{\rho}_k| \geq \hat{\rho}_k^{b,200,\text{corrected}}$ and then applying the simple deletion procedure (as described in Section 8.3.2)

- Setting 3 (CCTAD): Selection of all covariates $Z^{(k)}_{\text{duplicates}}$ with $|\hat{\rho}_k| \geq \hat{\rho}_k^{b,200,\text{corrected}}$ and then applying the advanced deletion procedure (as described in Section 8.3.3)

The simulation results are listed in Table 5. First of all, we can see that (CCT) selected

Covariate Selection	Cov	Bias	SD	Avg. SE	CI Length	Coverage
(CCT)	22.1350	-	-	-	-	-
(CCTSD)	2.2132	0.0066	0.0397	0.0377	0.1662	95.6
(CCTAD)	2.2132	0.0066	0.0397	0.0377	0.1662	95.6

Table 5: Simulation results using `RDHonest` with a sample size of 1,000 based on 10,000 Monte Carlo replications when including duplicated covariates

around 22 covariates. Since this set of selected covariates contained duplicates, the simulation ran into invertibility errors for the covariate adjustment. Therefore, this selection procedure delivers no estimation results here. As a result, we can conclude that (CCT) seems to not work appropriately in cases with duplicated covariates. This also seems to suggest that situations with a high number of pairwise correlated covariates might be problematic, leading to invertibility errors.

However, the simple and advanced deletion procedures seem to work as desired. Both reduce the number of selected covariates by (CCT) by a factor of approximately 10. This is exactly the result we would expect since every covariate is contained 10 times, which means that every favorable covariate was chosen 10 times by (CCT). This behaviour is verified by the results contained in Table 6, which shows the percentage of each covariate getting chosen by the respective selection procedures. Indeed, (CCT) chose the best covariates out of the 20 different covariates, but in turn included each of them 10 times. However, after the simple and advanced deletion all duplicates are removed and each covariate is just included once.

Overall, the result seems to suggest that (CCTSD) and (CCTAD) are superior to (CCT), particularly in settings with duplicated or correlated covariates.

9.4 Empirical Application

[CCW07] provides a data set on the Austrian labor market. For this reason, we refer to [CCW07] for a detailed description of its origin and construction. We want to re-conduct one of their analysis, which was also performed by [KR21], but this time by using our correlation-driven selection procedures.

During the time period covered by the data set, Austrian workers are entitled to receive a severance payment after losing their job if they worked for at least 36 month in the job by the time of separation. The scope of the analysis is to examine whether receiving a severance payment has an impact on the increase in the workers salary when comparing the salary in the new job to the one in the old job. This can indeed be addressed with an RDD analysis as the score is the duration of the old job, the treatment is receiving a severance payment and the outcome is the growth of the salary. Whenever the score is

Covariate	(CCT)	(CCTAD)	(CCTSD)
Number 1	100.00	99.06	99.06
Number 2	98.86	98.85	98.85
Number 3	20.78	20.77	20.77
Number 4	1.47	1.46	1.46
Number 5	0.15	0.15	0.15
Number 6	0.07	0.07	0.07
Number 7	0.02	0.02	0.02
Number 8 to 20	0.00	0.00	0.00
Number 21	100.00	0.00	0.00
Number 22	98.86	0.00	0.00
Number 23	20.78	0.00	0.00
Number 24	1.47	0.00	0.00
Number 25	0.15	0.00	0.00
Number 26	0.07	0.00	0.00
Number 27	0.02	0.00	0.00
Number 28 to 40	0.00	0.00	0.00
Number 41 to 60	Each as number 21 to 40		
Number 61 to 80			
Number 81 to 100			
Number 101 to 120			
Number 121 to 140			
Number 141 to 160			
Number 161 to 180			
Number 181 to 200			

Table 6: Percentages of the covariates being chosen by the different selection procedures in a scenario with duplicated covariates

greater or equal the threshold of 36 month, the treatment is applied. Also, the data set provides additional information which can be incorporated into the analysis as covariates. In fact, we have the following 60 covariates as considered in [KR21]:

- Basic covariates (38): Gender, marital status, Austrian nationality, "blue collar" occupation, age and its square, log of previous wage and its square, indicators for month and year of job termination
- Extended covariates (22): Work experience and its square, number of employees in the company where job was lost, indicator of having a job before the one just lost, "blue collar" status at job prior to the one lost, indicator of having a prior spell of non-employment, duration of last non-employment, total number of spells of non-employment in career, indicator of being recalled to the job before the one just lost, indicator for higher education, indicators for industry sector and region

Moreover, we want to incorporate all non-trivial interaction terms as well as all trigonometric transformations up to order five applied on all non-dummy covariates, i.e. $\sin(2\pi k \cdot Z)$ and $\cos(2\pi k \cdot Z)$ for $k = 1, ..., 5$ and all covariates Z. This leads to a total of $1,593$ covariates. Then, we perform RD analysis for the following scenarios:

- Setting 1 (CCT): Selection of all covariates $Z^{(k)}$ with $|\hat{\rho}_k| \geq \hat{\rho}_k^{b,1593,\text{corrected}}$ (as described in Section 8.3.1)
- Setting 2 (CCTSD): Selection of all covariates $Z^{(k)}$ with $|\hat{\rho}_k| \geq \hat{\rho}_k^{b,1593,\text{corrected}}$ and then applying the simple deletion procedure (as described in Section 8.3.2)
- Setting 3 (CCTAD): Selection of all covariates $Z^{(k)}$ with $|\hat{\rho}_k| \geq \hat{\rho}_k^{b,1593,\text{corrected}}$ and then applying the advanced deletion procedure (as described in Section 8.3.3)
- Setting 4: Selection of all covariates $Z^{(k)}$ with $|\hat{\rho}_k| \geq 0.2$
- Setting 5: Selection of no covariates
- Setting 6: Selection of all basic covariates
- Setting 7: Selection of all basic and extended covariates (This setting required to delete 3 covariates in order to obtain numerical invertibility.)

In the implementation, we remove each data entry having no value for at least one of the covariates, the score or the outcome. This leads to a data set containing 288,175 entries in total. However, a limitation of the data set to 100,000 entries was necessary due to limited RAM capacities. Therefore, the simulation was conducted using `RDHonest` on a data set containing 100,000 entries that were chosen randomly.

As we deal with a real data set here, and not with a generated data set as above, the true value of the average treatment effect is unknown. Therefore, we cannot provide the true bias and standard deviation as well as the coverage here. But, the following values will be given as outputs of our simulation results:

- Number of chosen covariates by the respective selection procedures,
- Estimated value of the average treatment effect (provided by `RDHonest`),
- Average standard error (provided by `RDHonest`)
- Lower and upper bound of the confidence interval as well as its length (provided by `RDHonest`)

Simulation results The simulation results are given in Table 7. Most importantly, we can see from the results that (CCT) chooses 310 covariates. This leads to invertibility issues during the covariate adjustment which is just reparable by deleting more than 200 of those covariates. But this would not represent the actual result of the selection anymore. However, the results are not of great interest anyways since we have seen in the above simulations that such a large number of included covariates tends to lead to a strongly downward biased standard error.

Besides that, all other estimates are approximately zero which coincides with the results stated in [CCW07] and [KR21]. We can see that (CCTSD) and (CCTAD) just select

Covariate Selection	Cov	Estimator	SE	CI Lower	CI Upper	CI Length
(CCT)	310	-	-	-	-	-
(CCTSD)	2	-0.0088	0.0171	-0.0455	0.0278	0.0733
(CCTAD)	22	-0.0125	0.0167	-0.0483	0.0234	0.0718
Corr. threshold of 0.2	5	-0.0190	0.0159	-0.0531	0.0150	0.0681
Fixed: no cov.	0	-0.0073	0.0197	-0.0498	0.0351	0.0849
Fixed: Basic	38	-0.0165	0.0153	-0.0495	0.0164	0.0659
Fixed: Basic+Ext.	57	-0.0101	0.0147	-0.0416	0.0215	0.0632

Table 7: Simulation results using `RDHonest` with the data set on the Austrian labor market

a small number of covariates, which indeed means that the vast majority of those 310 covariates selected by (CCT) were correlated to at least one other covariate contained in the selection. In particular, (CCTSD) selects just 2 covariates, while (CCTAD) selects 22 covariates.

Compared to the estimation without covariates, (CCTSD) improves the standard error by around 13 percent. Therefore, this selection procedure improves the estimation accuracy. The confidence interval is also shorter by around the same percentage. Since there are only two covariates selected, there is no reason for concern regarding a downward biased standard error here.

(CCTAD) improves the standard error even more by approximately 15 percent compared to the baseline estimator without covariates. The confidence interval indeed also gets smaller by around the same factor. This shows that the simple deletion procedure might drop some potentially relevant covariates. Furthermore, (CCTAD) only selects 22 covariates, which is still a small number in proportion to the data size of 100,000. In the above simulation with generated sample data of a size of 10,000, we can see that 30 covariates did not lead to an average standard error which significantly underestimates the true standard deviation. Therefore, especially here with a sample size of 100,000, there seems to be no reason for concern about a downward biased standard error. Hence, the results seem to suggest that (CCTAD) is superior to (CCTSD) as it delivers even better estimation results while selecting an appropriate number of covariates in relation to the sample size. In fact, (CCTSD) seems to miss out on potentially useful covariates.

In that regard, we also want to list the chosen covariates by (CCTAD) and (CCTSD). In fact, (CCTSD) selected the covariates

- Wage in previous job
- Interaction term: Square of wage in previous job and indicator for month of job termination

This seems reasonable since the wage in the previous job is intuitively the most related information to the effect we want to estimate. (CCTAD) selected the covariates

- Wage in previous job

- Indicator for year of job termination
- Interaction term: Age and "blue collar" occupation
- Interaction term: Square of wage in previous job and number of employees in the company where job was lost
- Interaction term: Square of wage in previous job and indicator of having a prior spell of non-employment
- Interaction term: Age and indicator for year of job termination
- Interaction terms (in total 11): Square of wage in previous job and indicators for month and year of job termination
- Interaction terms (in total 2): Square of wage in previous job and indicators for region
- Interaction term: Square of work experience and indicator for month of job termination
- Interaction term: Duration of last non-employment and indicator for month of job termination
- Interaction term: Indicator for region and indicator for year of job termination

which indeed also suggests that mostly basic covariates are taken into account. However, the interaction terms also contain some covariates of the extended set.

At last, we want to mention that the fixed correlation threshold of 0.2 again seems to be an appropriate choice as it selects a reasonable number of 5 covariates.

10 Conclusion

Under sufficient regularity conditions, the local linear covariate-adjusted estimator of the average treatment effect including a fixed set of additional covariates is asymptotically normally distributed. The conducted simulations confirm that a small number of well-selected covariates can increase the estimation accuracy. However, we have also seen that including too many covariates leads to a progressively strong downward bias in the standard error, causing too small confidence intervals. For this reason, the correlation-based selection procedure (CCT) does not seem to be an appropriate approach as the number of selected covariates can get quiet large. On the contrary, the two selection procedures (CCTSD) and (CCTAD) via calculated correlation threshold with, respectively, simple and advanced deletion seem to choose a rather small set of covariates which is highly related to the outcome. This suggests that these methods can be used to select covariates which meaningfully improve estimation accuracy.

10.1 Future Work

[KR21] provides conditions which a general selection algorithm has to satisfy in order that its post-selection estimator is asymptotically normally distributed. As the two selection procedures (CCTSD) and (CCTAD) delivered favorable results in the simulations conducted in this thesis, one could examine whether they fulfill those stated conditions.

10.2 Acknowledgements

I would like to thank Jun.-Prof. Dr. Alexander Kreiß for his guidance and extremely helpful advice. I am very grateful for the time he spent to discuss all of my questions as well as the provided knowledge and expertise. Also, I appreciate his comments on my work in progress.

References

[AK18] Timothy B. Armstrong and Michal Kolesár. Optimal inference in a class of regression models. *Econometrica*, 86(2):655–683, 2018.

[AOS21] Yoichi Arai, Taisuke Otsu, and Myung Hwan Seo. Regression Discontinuity Design with Potentially Many Covariates, 2021.

[CCFT19] Sebastian Calonico, Matias D. Cattaneo, Max H. Farrell, and Rocio Titiunik. Regression Discontinuity Designs Using Covariates. *The Review of Economics and Statistics*, 2019.

[CCFT22] Sebastian Calonico, Matias D. Cattaneo, Max H. Farrell, and Rocio Titiunik. *rdrobust: Robust Data-Driven Statistical Inference in Regression-Discontinuity Designs*, 2022. R package version 2.1.1. `https://CRAN.R-project.org/package=rdrobust`. Last accessed 8 April 2023.

[CCT14] Sebastian Calonico, Matias D. Cattaneo, and Rocio Titiunik. Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs. *Econometrica*, 2014.

[CCW07] David Card, Raj Chetty, and Andrea Weber. Cash-on-Hand and Competing Models of Intertemporal Behavior: New Evidence from the Labor Market. *The Quarterly Journal of Economics*, 122(4):1511–1560, 2007.

[CIT19] Matias D. Cattaneo, Nicolas Idrobo, and Rocio Titiunik. A Practical Introduction to Regression Discontinuity Designs: Foundations. *Cambridge Elements: Quantitative and Computational Methods for Social Science*, 2019.

[CIT23] Matias D. Cattaneo, Nicolas Idrobo, and Rocio Titiunik. A Practical Introduction to Regression Discontinuity Designs: Extensions. *Cambridge Elements: Quantitative and Computational Methods for Social Science*, 2023.

[CT21] Matias D. Cattaneo and Rocio Titiunik. Regression Discontinuity Designs. `https://conference.nber.org/confer/2021/SI2021/ML/RDD.pdf`, 2021. Last accessed 29 January 2023.

[CTVB17] Matias D. Cattaneo, Rocio Titiunik, and Gonzalo Vazquez-Bare. Comparing Inference Approaches for RD Designs: a Reexamination of the Effect of Head Start on Child Mortality. *Journal of Policy Analysis and Management*, 36:643–681, 2017.

[CTVB20] Matias D. Cattaneo, Rocio Titiunik, and Gonzalo Vazquez-Bare. Analysis of regression-discontinuity designs with multiple cutoffs or multiple scores. *The Stata Journal*, 2020.

[Dur19] Rick Durrett. *Probability: Theory and Examples*, page 29. Cambridge University Press, 5th edition, 2019.

[FH19] Markus Frölich and Martin Huber. Including Covariates in the Regression Discontinuity Design. *Journal of Business & Economic Statistics*, 37(4):736–748, 2019.

[HTdK01] Jinyong Hahn, Petra Todd, and Wilbert Van der Klaauw. Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica*, 69(1):201–209, 2001.

[IK12] Guido Imbens and Karthik Kalyanaraman. Optimal Bandwidth Choice for the Regression Discontinuity Estimator. *The Review of Economic Studies*, 2012.

[IL08] Guido W. Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635, 2008. The regression discontinuity design: Theory and applications.

[Kle14] Achim Klenke. *Probability Theory*, pages 357–363. Springer Berlin Heidelberg, 2nd edition, 2014.

[Kol23] Michal Kolesár. *RDHonest: Honest Inference in Regression Discontinuity Designs*, 2023. R package version 0.4.1. `https://github.com/kolesarm/RDHonest`. Last accessed 8 April 2023.

[KR21] Alexander Kreiß and Christoph Rothe. Inference in Regression Discontinuity Designs with High-Dimensional Covariates, 2021.

[LL10] David S. Lee and Thomas Lemieux. Regression Discontinuity Designs in Economics. *Journal of Economic Literature*, 48(2):281–355, June 2010.

[NOR21] Claudia Noack, Tomasz Olma, and Christoph Rothe. Flexible Covariate Adjustments in Regression Discontinuity Designs, 2021.

[Wic16] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. `https://ggplot2.tidyverse.org`. Last accessed 8 April 2023.

[ZZW03] Donald W. Zimmerman, Bruno D. Zumbo, and Richard H. Williams. Bias in Estimation and Hypothesis Testing of Correlation. *Psicológica*, 2003.

Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe, insbesondere sind wörtliche oder sinngemäße Zitate als solche gekennzeichnet. Mir ist bekannt, dass Zuwiderhandlung auch nachträglich zur Aberkennung des Abschlusses führen kann.
Ich versichere, dass das elektronische Exemplar mit den gedruckten Exemplaren übereinstimmt.

Leipzig, den 18.04.2023

.....
Patrick Kramer