A Conditional Moment Based Approach for Partially Linear Models with Applications in Treatment Effects

by

Xiaolin Sun

M.A., Simon Fraser University, 2016
B.A., China University of Political Science and Law, 2013

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Economics
Faculty of Arts and Social Sciences

SIMON FRASER UNIVERSITY
Summer 2022

Declaration of Committee

Name:	**Xiaolin Sun**
Degree:	**Doctor of Philosophy**
Thesis title:	**A Conditional Moment Based Approach for Partially Linear Models with Applications in Treatment Effects**
Committee:	**Chair:** Simon Woodcock Associate Professor, Economics
	Bertille Antoine Supervisor Professor, Economics
	Krishna Pendakur Committee Member Professor, Economics
	Dongwoo Kim Committee Member Assistant Professor, Economics
	Kevin Schnepel Examiner Assistant Professor, Economics
	Vadim Marmer External Examiner Professor, Vancouver School of Economics University of British Columbia

Abstract

In the first chapter of the thesis, we propose a new estimator for the slope parameter of the endogenous variable of interest in a partially linear conditional moment model, which combines a Robinson transformation (Robinson (1988)), to partial out the non-linear part of the model with a smooth minimum distance (SMD) approach (Lavergne and Patilea (2013)), to exploit all the information of the conditional mean independence restriction. Our estimator only depends on one tuning parameter, is easy to compute, consistent and $\sqrt{n}$-asymptotically normal under standard regularity conditions. Simulations show that our estimator is competitive with GMM-type estimators, and often displays a smaller bias and variance, as well as better coverage rates for confidence intervals. We revisit and extend some of the empirical results in Dinkelman (2011) who estimates the impact of electrification on employment growth in South Africa. Overall, we obtain estimates that are smaller in magnitude, more precise, and still economically relevant.

In the second chapter, we develop a new estimator for heterogeneous treatment effects in a partially linear model (PLM) with endogenous treatment. The PLM has a parametric part that includes the treatment and the interactions between the treatment and exogenous characteristics, and a nonparametric part that contains those characteristics and many other covariates. The new estimator is a combination of the estimator proposed in the first chapter and a Neyman-Orthogonalized first-order condition (NOFOC). Our estimator, using only one valid binary instrument, identifies both parameters. With the sparsity assumption, using regularised machine learning methods (i.e., the Lasso method) allows us to choose a relatively small number of polynomials of covariates. Our new estimator is less biased, consistent, and $\sqrt{n}$-asymptotically normal under standard regularity conditions. Simulations show that our estimator behaves well with different sets of instruments, but the GMM-type estimators do not. We use the Card application to show the differences between estimators using various sets of instruments. It shows that our new method generates more precise estimates in comparison to GMM.

In the third chapter, we estimate the heterogeneous treatment effects of Medicaid on individual outcome variables from the Oregon Health Insurance Experiment. In this experiment, our method from the previous chapter produces more significant and more reliable results for heterogeneous effects of health coverage on economic outcomes.

Keywords: Robinson's transformation; Conditional mean independence; Non-linearity; Minimum distance estimation; regularised machine learning method; Instrument

Dedication

This thesis is dedicated to my parents, who have always been extremely supportive.

Acknowledgements

I would like to express my deepest gratitude to Bertille Antoine, my supervisor, for her continuous encouragement, unwavering support, and generosity. Her advice and suggestions aided me in finishing my thesis.

Krishna Pendakur and Dongwoo Kim have also provided valuable suggestions and remarks. I also very much appreciate the guidance and comments from them.

I would like to convey my gratefulness to Chris Bidner, Irene Botosaru, Brian Krauth, Chris Muris, Kevin Schnepel, Hitoshi Shigeoka, Xiaoting Sun, and Simon Woodcock for their insightful advice and feedback.

Table of Contents

List of Tables

List of Figures

Chapter 1

Partially Linear Models with Endogeneity: a conditional moment based approach

1 Introduction

Many empirical studies focus on estimating the causal or structural effect of some variable on an outcome of interest: for example, Dinkelman (2011) is interested in estimating the impact of electrification on employment growth in South Africa. Since policies and many other variables of interest are not exogenous, researchers often rely on an instrumental variable - assumed to be valid and relevant - after controlling for a sufficient set of other factors or covariates. More specifically, these studies are often based on a first stage linear in the instrument and on a second stage linear in the covariates. Using only linear first and second stages may miss important information about effect heterogeneity as well as instrument and covariate validity.

We provide a framework that extends the above-mentioned standard linear setup in two main directions: (i) we consider a partially linear model to allow (exogenous) covariates Z to enter non-parametrically in the second stage; (ii) we bypass the parametrization of the first stage and directly rely on the informational content of the maintained conditional mean independence of the instrument W to estimate the parameters of the parametric part of the model (e.g. the slope parameter of the endogenous variable). More specifically, our estimation procedure combines Robinson's transformation (Robinson (1988)) with a smooth minimum distance approach (SMD, Lavergne and Patilea (2013)). The Robinson's transformation partials out the non-parametric part of the model which is seen as a nuisance parameter, while the SMD estimation conveniently exploits all the information of the conditional mean independence restriction without having to parametrically model or estimate the first stage. Our estimator only depends on one tuning parameter, is easy to compute and available in closed form in our partially linear framework. It is consistent and $\sqrt{n}$-asymptotically normal under standard regularity conditions. Simulations show that our estimator is competitive with various GMM-type estimators, and often displays a smaller bias and variance, as well as better coverage rates for confidence intervals. To illustrate the performance of our estimator in practice, we revisit and extend some of the empirical results in Dinkelman (2011) who estimates the impact of electrification on employment growth in South Africa: overall, we obtain estimates that are smaller in magnitude, but more precise. In particular, we report statistically and economically significant effect of electrification on all measures of household energy sources and household services, on population growth, and on the change in fraction of women that have a completed high school education, but no effect on employment rate.

Our work contributes to the literature on partially linear models which has a long tradition in economics and statistics: see Engle et al. (1986) for an early application and the monograph by Härdle et al. (2000). More broadly, our model belongs to the general class of conditional moment restriction models with unknown functions considered by Ai and Chen (2003) which allows for the presence of endogeneity in the parametric and nonparametric parts of the model. As an alternative to the estimation strategy proposed in this paper, one may consider combining K (unconditional) moments derived from the conditional mean independence using a criterion function like the generalized method of moments (GMM) after approximating the nonparametric part of the model through L basis functions like power series or splines. For example, Ai and Chen (2003) rely on a sieve minimum distance (which can be interpreted as GMM-type procedure), while Otsu (2011) puts forward a sieve conditional empirical likelihood approach. Alternative asymptotics may require L and K to grow with the sample size: see Cattaneo et al. (2018a) and references therein for results obtained with "many covariates"; see Bekker (1994), Chao et al. (2012), Hausman et al. (2012) for results obtained with "many instruments". In this context, it is important to work with an increasing number K of (unconditional) moments to obtain an equivalent information set; see e.g. Dominguez and Lobato (2004) for some examples and related discussion. Further, for consistency, K cannot grow too fast with respect to the sample size. Therefore, in applications, the chosen number of instruments is likely to influence empirical results, and its selection appears to be an important, but delicate task. By directly exploiting the informational content of the conditional mean independence (without having to estimate it), we do not need to choose the number of instruments, or a smoothing parameter associated with first-stage estimation[1] that may affect consistency. Our estimation procedure also avoids having to choose how the instruments enter the selected (unconditional) moments (say $h(W)$ with $h(.)$ chosen vector of K measurable functions), which has been shown to affect the reliability of the estimator; see e.g. Jun and Pinkse (2012).

[1] These properties are shared by the estimators proposed by Dominguez and Lobato (2004) and Lavergne and Patilea (2013). Neither consider partially linear models.

In order to partial out the non-parametric part of the model by applying Robinson's transformation, we do need consistent estimation of conditional expectations (taken with respect to the exogenous covariates Z, say $E(.|Z)$). Accordingly, our procedure depends on one tuning parameter: specifically, the bandwidth value for the associated Nadaraya-Watson estimator, which is straightforward to choose in practice (for example, using cross-validation). As already mentioned, other non-parametric methods could be considered instead, and would then depend on choosing L (e.g. the number of neighbors when using nearest-neighbors, or the number of terms when using series-based estimation): it is less clear in practice how to choose such a tuning parameter[2].

One may be able to do without such a consistent (kernel-based) estimator, or, at least, mitigate its impact by exploiting recent results obtained by Cattaneo et al. (2018b) with series-based estimation, or by Chernozhukov et al. (2018) with (double) machine-learning. We leave these investigations for future work.

Finally, our work contributes to a recent literature that highlights some shortcomings of the traditional two-stage least squares procedure that relies on a linear first stage obtained by regressing the endogenous variable on the instrument: see e.g. Xu (2019) and references therein with a binary (endogenous) variable. Our framework allows for a fully flexible first stage without having to parametrically model or estimate it, but maintains a (standard) strong identification assumption: for extensions beyond strong identification, see Antoine and Lavergne (2020) who link identification issues to a (linear) first stage that does not appropriately capture the variation of the endogenous variable.

Our paper is organized as follows. In section 2, we introduce our framework and motivate our estimation strategy. The asymptotic properties of our estimator are derived in section 3. We illustrate its finite sample properties through a Monte-Carlo study in section 4 and by revisiting some of the empirical results in Dinkelman (2011) in section 5. The graphs and tables of Monte-Carlo and empirical results are collected in the Appendix. The proofs of our theoretical results, as well as additional theoretical and empirical results, are presented in the supplementary Appendix.

2 Framework and Motivation

We consider partially linear models with endogeneity in the parametric part,

$$y_i = X_i' \beta_0 + g(Z_i) + e_i \tag{2.1}$$

where the dependent variable y_i is scalar, X_i is the vector of p explanatory variables, β_0 is the unknown vector of p parameters of interest, and $g(.)$ is an unknown (sufficiently

[2] Recent results in Breunig and Chen (2020) may alleviate some of these concerns.

smooth) function of q_z exogenous variables Z_i. We are interested in estimating β_0 - but not necessarily $g(.)$. In order to do so, we rely on a vector W_i of $q_w \geq p$ instruments that may include components of X_i (when their exogeneity is maintained), Z_i, and some additional variables, such that

$$E(e_i|W_i) = 0 \ a.s. \tag{2.2}$$

The data is assumed to be i.i.d. and we allow for a conditionally heteroskedastic error process of unknown form, $E(e_i^2|w) = \sigma^2(w)$.

Since e_i depends on the unknown function $g(.)$, we cannot directly use the conditional moment restriction (2.2): to circumvent this difficulty, we first apply a Robinson's transformation (Robinson (1988)) to the original model. Under the maintained exogeneity assumption of Z_i, this amounts to subtracting the conditional expectation of y_i with respect to Z_i from (2.1) to get

$$y_i - E(y_i|Z_i) = (X_i - E(X_i|Z_i))'\beta_0 + e_i \qquad \text{with} \qquad E(e_i|W_i) = 0 \ a.s.$$

and W_i includes Z_i. Using obvious notations, this can be rewritten as

$$\tilde{y}_i = \tilde{X}_i'\beta_0 + e_i \qquad \text{where} \qquad E(\tilde{y}_i - \tilde{X}_i'\beta_0|W_i) = 0 \ a.s. \tag{2.3}$$

In the traditional GMM setting, a finite number of unconditional moments is then extracted from (2.3) by considering instruments taken as (measurable) functions of W_i: e.g. simply W_i in the traditional 2SLS approach. A large amount of information is discarded by doing so as explained in Dominguez and Lobato (2004), but different functions of the instrument will only affect efficiency, as they should all identify the same population parameter under the classic framework of "homogeneous effect"[3]. Since there is often little to no information on the relationship between endogenous variable and instrument, an estimation strategy that leaves the functional form of the first stage equation unspecified should be valuable for empirical analysis. To directly use the informational content of the conditional moment restriction (2.3) without having to rely on its parametrization, we follow an original idea by Bierens (1982) and rewrite (2.3) as an (equivalent) continuum of unconditional moment restrictions:

$$E\left[(\tilde{y}_j - \tilde{X}_j'\beta_0)e^{it'W_j}\right] = 0 \quad \forall t \in \mathbb{R}^{q_w} \tag{2.4}$$

[3] Sensitivity to the first stage - e.g. as documented in Dieterle and Snell (2016) - signals an invalid instrument, or unmodeled heterogeneity in the sense that different first stages identify different weighted averages of underlying responses; see e.g. Angrist et al. (2000) and Heckman et al. (2006).

The main idea is thus to build a theoretical criterion that combines the above continuum of restrictions into a single criterion, uniquely minimized at β_0. The Integrated Conditional Moment (ICM) principle (Bierens (1982)) replaces conditional moment restrictions by a continuum of unconditional moments such as (2.4). Other functions have been used beyond the complex exponential, see Bierens (1990) and Bierens and Ploberger (1997). Stinchcombe and White (1998) give a characterization of a large class of functions that could generate an equivalent set of unconditional moments. As detailed by Lavergne and Patilea (2013), this yields a full collection of potential estimators such as the ones developed by Dominguez and Lobato (2004), Antoine and Lavergne (2014), or Escanciano (2018) among others. We focus here on a particular application of the ICM suitable for theoretical investigation and practical implementation, and we leave for future work the investigation of the relative merits of these different ICM-type estimators.

For a given strictly positive measure μ on $\mathbb{R}^{q_w}$, our theoretical objective function is defined as

$$\begin{aligned} M_\infty(\beta) &= \int_{\mathbb{R}^{q_w}} |E(e_j(\beta)e^{it'W_j})|^2 d\mu(t) \qquad (2.5) \\ \text{where} \qquad e_j(\beta) &\equiv y_j - E(y_j|Z_j) - (X_j - E(X_j|Z_j))'\beta \end{aligned}$$

The minimization of the objective function defined in (2.5) can only be solved numerically. We consider instead the following population objective function after introducing the Fourier transform $k(.)$ of the density induced by μ,

$$\begin{aligned} M_\infty(\beta) &= E\left[e_j(\beta)e_l(\beta)\kappa_{j,l}\right] \qquad (2.6) \\ \text{where} \qquad \kappa_{j,l} &\equiv k(W_j - W_l) = \int_{\mathbb{R}^{q_w}} e^{it'(W_j - W_l)} d\mu(t) \quad \forall j \neq l \end{aligned}$$

and (y_l, X_l, W_l) an independent copy of (y_j, X_j, W_j). The definitions of $M_\infty(.)$ in (2.5) and (2.6) are the same[4]. With the alternative population objective function (2.6), we avoid calculating the derivative of the norm of a complex function. In addition, we show in Proposition 1 below that (2.6) is uniquely minimized at β_0 (under some regularity conditions) and derive a closed-form expression for β_0. Our first set of regularity assumptions is presented next.

Assumption 1. *(Regularity assumptions)*
(i) $E(e_j|W_j) = 0$ *and* W_j *includes* Z_j*.*
(ii) $E(\tilde{X}_j|W_j) \neq 0$ *a.s. (with probability 1) with* $\tilde{X}_j = X_j - E(X_j|Z_j)$*.*
(iii) $E(\tilde{X}_j\tilde{X}_j')$ *is nonsingular.*
(iv) Let $f_W(.)$ *denote the density function of* W_j*. We assume that* $E(\tilde{X}_j|W_j = .)f_W(.)$ *is* L_q *for some* $1 \leq q \leq 2$*.*
(v) (y_l, X_l, W_l) *is an independent copy of* (y_j, X_j, W_j)*.*

[4] See the Supplementary Appendix for a formal proof that (2.5) and (2.6) coincide.

(vi) Let μ be a given strictly positive measure on $\mathbb{R}^{q_w}$. Let $k(.)$ be the Fourier transform induced by μ, $k(W_j - W_l) = \int_{\mathbb{R}^{q_w}} e^{it'(W_j - W_l)} d\mu(t)$. We assume that $k(.)$ is a symmetric bounded density function on $\mathbb{R}^{q_w}$ and that its Fourier transform is strictly positive.

Assumption $1(i)$ maintains the validity of the instruments W_j and the exogeneity of Z_j, while Assumption $1(ii)$ maintains the relevance (and strength) of W_j needed to identify β_0. It also implies that there exists a measurable function $f(.)$ such that $E(\tilde{X}_j f(W_j)) \neq 0$. Specifically, there exists t such that $E(\tilde{X}_j e^{it'W_j}) \neq 0$. Assumption $1(iii)$ is the same identification assumption as the one maintained by Robinson (1988) (see his condition (3.5)). Assumption $1(iv)$ is mild and sufficient to ensure existence of the corresponding Fourier transform. Assumption $1(vi)$ is not too restrictive on the measure μ (and associated $k(.)$). Examples of suitable densities include products of triangular, normal, logistic (see Johnson et al. (1995), Section 23.3), Student[5] (including Cauchy, see Hurst (1995)), or Laplace densities.

Proposition 1. *(Identification of β_0)*
Under Assumption 1, β_0 is the unique minimizer of (2.6) with $M_\infty(\beta_0) = 0$ and

$$\beta_0 = \left[E(\kappa_{j,l}\tilde{X}_j\tilde{X}_l')\right]^{-1} E(\kappa_{j,l}\tilde{X}_j\tilde{y}_l)$$

where $\tilde{y}_j \equiv y_j - E(y_j|Z_j)$ and $\tilde{X}_j \equiv X_j - E(X_j|Z_j)$.

A natural estimator of β_0 minimizes a sample analog of (2.6) obtained after replacing the expectation by a double average. Therefore, an (infeasible) estimator of β_0 is defined as

$$\tilde{\beta}_n = \arg\min_{\beta \in B} M_n(\beta) \qquad (2.7)$$

$$\text{with} \qquad M_n(\beta) = \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{l \neq j}^{n} e_j(\beta) e_l(\beta) k(W_j - W_l) \qquad (2.8)$$

This is a special case of the Smooth Minimum Distance (SMD) estimator introduced by Lavergne and Patilea (2013) when a fixed bandwidth (equal to 1) is used: they show that it is consistent and asymptotically normally distributed in a framework with general conditional estimating equations. In our linear framework, it is available in closed-form,

$$\tilde{\beta}_n = [\tilde{X}'\tilde{\kappa}\tilde{X}]^{-1}\tilde{X}'\tilde{\kappa}\tilde{y}\,,$$

[5] Student density should be chosen with enough degrees of freedom to ensure other regularity conditions are satisfied.

with $\tilde{y}$ the $(n,1)$-vector with elements $\tilde{y}_j$, $\tilde{X}$ the (n,p)-matrix with rows $\tilde{X}_j'$, and $\tilde{\kappa}$ the (n,n)-matrix with (j,l) element $\kappa_{j,l}$ (when $j \neq l$) and 0 on the main diagonal[6].

In the next section, we present the asymptotic properties of $\tilde{\beta}_n$, and introduce our feasible estimator that shares its asymptotic properties.

3 Large sample theory of R-SMD

In this section, we first present the asymptotic properties of $\tilde{\beta}_n$, the infeasible SMD estimator of β_0. Then, we introduce our Robinson-SMD (R-SMD hereafter) estimator, as a feasible estimator that shares the same asymptotic properties as $\tilde{\beta}_n$.

3.1 Asymptotic properties of the infeasible estimator $\tilde{\beta}_n$

The infeasible estimator $\tilde{\beta}_n$ is a special case of the SMD estimator introduced by Lavergne and Patilea (2013), and its asymptotic properties are known (see their Theorems 2.1 and 2.2). We present these results in our simpler linear framework with a fixed bandwidth.

Assumption 2. *(Regularity assumptions)*
(i) We consider a sample of n iid observations (y_j, X_j, W_j) for $j = 1, \cdots, n$.
(ii) Let $f_W(.)$ denote the density function of W_j. Let $\tilde{X}_{j,r}$ denote the r-th component of $\tilde{X}_j$ with $1 \leq r \leq p$. We assume that $E(e_j^2|W_j = .)f_W(.)$, $E(\tilde{X}_{j,r}e_j|W_j = .)f_W(.)$, and $E(\tilde{X}_{j,r}\tilde{X}_{j,s}|W_j = .)f_W(.)$ are L_q for some $1 \leq q \leq 2$ for any r and s such that $1 \leq r, s \leq p$.

Assumption 2(*ii*) maintains sufficient conditions to ensure the applicability of central limit theorems for appropriate U-statistics.

Proposition 2. *(Consistency and Asymptotic normality of $\tilde{\beta}_n$)*
Our infeasible estimator $\tilde{\beta}_n$ is defined as

$$\tilde{\beta}_n \equiv \arg\min_{\beta} M_n(\beta) = [\tilde{X}'\tilde{\kappa}\tilde{X}]^{-1}\tilde{X}'\tilde{\kappa}\tilde{y}\,.$$

Under Assumptions 1 and 2, $\tilde{\beta}_n$ is consistent for β_0, that is $\tilde{\beta}_n \xrightarrow{p} \beta_0$, and asymptotically normally distributed,

$$\sqrt{n}(\tilde{\beta}_n - \beta_0) \xrightarrow{d} \mathcal{N}\left(0, \left[E(\kappa_{j,l}\tilde{X}_j\tilde{X}_l')\right]^{-1} \texttt{Var}\left[h_1(\tilde{X}_j, e_j(\beta_0), W_j)\right] \left[E(\kappa_{j,l}\tilde{X}_j\tilde{X}_l')\right]^{-1}\right)$$

$$\textit{with} \quad \texttt{Var}\left[h_1(\tilde{X}_j, e_j(\beta_0), W_j)\right] \equiv \texttt{Var}\left[\int_{\mathbb{R}^{q_w}} e^{-it'W_j} e_j(\beta_0) E[e^{it'W_l}\tilde{X}_l] d\mu(t)\right]$$

[6] The computation of the matrix $\tilde{\kappa}$ simplifies greatly for convenient choices of μ such as the standard normal; see the supplementary Appendix for additional results.

Our infeasible estimator relies on $n(n-1)$ pairs of observations, and its asymptotic properties will be derived using U-statistics: under Assumption 2, LLN and CLT hold for the appropriate U-statistics. The asymptotic variance of our infeasible estimator $\tilde{\beta}_n$ has a (traditional) sandwich form, but it is not efficient since we consider a fixed bandwidth; for a thorough discussion of the efficient SMD estimator, we refer the reader to section 2.5 in Lavergne and Patilea (2013) as well as to section 3.3 below. Note also that the asymptotic variance involves a complex integral in the middle term. However, it is important to mention that its imaginary part will vanish thanks to Assumption 1(vi): since $k(W_j - W_l)$ is a symmetric density function on $\mathbb{R}^{q_w}$, μ is symmetric as well; see also (3.10) below for a consistent estimator.

3.2 Feasible estimator of β_0

As already explained, the estimator $\tilde{\beta}_n$ depends on $\tilde{y}$ and $\tilde{X}$ that are unknown in practice due to the presence of the following conditional expectations $E(y_j|Z_j)$ and $E(X_j|Z_j)$. In order to propose a feasible estimator, these conditional expectations are replaced by their Nadaraya–Watson kernel estimators, respectively denoted $\hat{g}_y(Z_i)$ and $\hat{g}_X(Z_i)$. Our feasible R-SMD estimator $\hat{\beta}_n$ minimizes the feasible counterpart of $M_n(\beta)$ obtained after replacing the above-mentioned conditional expectations with their kernel estimators. We show that, under some regularity assumptions, our feasible R-SMD estimator $\hat{\beta}_n$ shares the same asymptotic properties as the infeasible SMD estimator $\tilde{\beta}_n$.

Our feasible R-SMD estimator $\hat{\beta}_n$ is defined as

$$\hat{\beta}_n \quad \equiv \quad \arg\min_{\beta \in B} \frac{1}{n(n-1)} [\widehat{\tilde{y}} - \widehat{\tilde{X}}\beta]' \tilde{\kappa} [\widehat{\tilde{y}} - \widehat{\tilde{X}}\beta] = [\widehat{\tilde{X}}{}' \tilde{\kappa} \widehat{\tilde{X}}]^{-1} \widehat{\tilde{X}}{}' \tilde{\kappa} \widehat{\tilde{y}} \tag{3.9}$$

$$\begin{aligned} \textit{with} \qquad & \widehat{\tilde{y}} \text{ the } (n,1)\text{-vector with elements } \widehat{\tilde{y}}_j \equiv y_j - \hat{g}_y(Z_j) = y_j - \frac{\sum_{i=1}^n y_i K(\frac{Z_i - Z_j}{h})}{\sum_{i=1}^n K(\frac{Z_i - Z_j}{h})}, \\ & \widehat{\tilde{X}} \text{ the } (n,p)\text{-matrix with rows } \widehat{\tilde{X}}{}'_j \equiv X'_j - \hat{g}'_X(Z_j) = X'_j - \frac{\sum_{i=1}^n X'_i K(\frac{Z_i - Z_j}{h})}{\sum_{i=1}^n K(\frac{Z_i - Z_j}{h})}, \end{aligned}$$

$\tilde{\kappa}$ the (n,n)-matrix with (j,l) element $\kappa_{j,l}$ (when $j \neq l$) and 0 on the main diagonal, and $K(.)$ a second-order product kernel and h a vanishing bandwidth as defined in Assumption 3 below.

In order to derive the asymptotic properties of $\hat{\beta}_n$, we need additional regularity assumptions.

Assumption 3. *(Regularity of $g(.)$ and of the kernel estimator)*
(i) The function $g(.)$ in our main model (2.1) is sufficiently smooth to be partialled out by the Robinson's tranformation.

(ii) $K(.)$ is a product kernel based on a second-order univariate kernel $k(.)$ such that

$$\begin{aligned} K\left(\frac{Z_i - Z_j}{h}\right) &= \Pi_{s=1}^{q_z} k\left(\frac{Z_{s,i} - Z_{s,j}}{h_s}\right) \\ \textit{with} \qquad h \equiv \Pi_{s=1}^{q_z} h_s \quad \textit{and} \quad & \sqrt{n}\left(\sum_{s=1}^{q_z} h_s^4 + \left[\frac{1}{nh_1 ... h_{q_z}}\right]\right) = o(1) \end{aligned}$$

Assumption 3(i) maintains similar regularity conditions on $g(.)$ as done in Robinson (1988) on p939. Assumption 3(ii) is used to control the bias of the non-parametric kernel estimator of $g_y(.)$ and $g_X(.)$. We usually assume that the bandwidth h converges to 0: under Assumption 3, h actually converges to 0 faster to ensure that the bias of the Nadaraya–Watson estimator converges to 0 as n increases. It is well-known (see e.g. Li and Racine (2007)) that, for Nadaraya–Watson kernel estimator, we have:

$$\begin{aligned} \hat{g}_y(Z_i) - E(y_i|Z_i) &= \mathcal{O}_p\left(v_n\right) \quad \text{and} \quad \hat{g}_X(Z_i) - E(X_i|Z_i) = \mathcal{O}_p\left(v_n\right) \\ \text{with} \quad v_n &\equiv \sum_{s=1}^{q_z} h_s^2 + \left[\frac{1}{nh_1 ... h_{q_z}}\right]^{0.5} \end{aligned}$$

Our next result presents the asymptotic properties of our R-SMD estimator.

Theorem 3.1. *(Consistency and Asymptotic normality of $\hat{\beta}_n$)*
Under Assumptions 1, 2, and 3, the R-SMD estimator $\hat{\beta}_n$ defined in (3.9) is consistent for β_0, that is $\hat{\beta}_n \xrightarrow{p} \beta_0$, and asymptotically normally distributed,

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} \mathcal{N}\left(0, \left[E(\kappa_{j,l}\tilde{X}_j\tilde{X}_l')\right]^{-1} \mathtt{Var}\left[h_1(\tilde{X}_j, e_j(\beta_0), W_j)\right] \left[E(\kappa_{j,l}\tilde{X}_j\tilde{X}_l')\right]^{-1}\right)$$

$$\textit{with} \quad \mathtt{Var}\left[h_1(\tilde{X}_j, e_j(\beta_0), W_j)\right] \equiv \mathtt{Var}\left[\int_{\mathbb{R}^{q_w}} e^{-it'W_j} e_j(\beta_0) E[e^{it'W_l}\tilde{X}_l] d\mu(t)\right]$$

Theorem 3.1 shows that our feasible R-SMD estimator $\hat{\beta}_n$ shares the same asymptotic properties as the infeasible estimator $\tilde{\beta}_n$. This follows, in part, from Assumption 3 which ensures that the mean squared error of the kernel estimation converges uniformly to zero at a rate faster than $\sqrt{n}$. This result generalizes Robinson's approach to a conditional moment framework à la Bierens (1982).

The chosen measure $\mu(.)$ does not affect the consistency or the rate of convergence of the R-SMD estimator, but it does affect its asymptotic variance. However, no measure $\mu(.)$ can be expected to deliver a better estimator (that is, an estimator with a smaller asymptotic variance) for every single underlying DGP. In our simulation study in section 4, we verify that the chosen $\mu(.)$ does not affect the performance of the R-SMD estimator much. It is also important to mention that the combination of the continuum of moments in our theoretical objective function, as well as in our estimator, is not optimal in general. Such an optimal combination is a difficult issue which is informally discussed in the next subsection.

The R-SMD estimator's asymptotic variance can be estimated in a standard heteroskedastic-robust way using an Eicker-White type approach,

$$\begin{aligned}
&\left[\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\widehat{\tilde{X}}_j\widehat{\tilde{X}}_l'\right]^{-1}\sum_{j=1}^{n}\left[\left(\sum_{l=1}^{n}\kappa_{j,l}\widehat{\tilde{X}}_l\right)\left(\sum_{l=1}^{n}\kappa_{j,l}\widehat{\tilde{X}}_l'\right)\hat{e}_j^2\right]\left[\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\widehat{\tilde{X}}_j\widehat{\tilde{X}}_l'\right]^{-1} \\
= \; &\left[\widehat{\tilde{X}}'\tilde{\kappa}\widehat{\tilde{X}}\right]^{-1}\widehat{\tilde{X}}'\tilde{\kappa}\Omega_n\tilde{\kappa}\widehat{\tilde{X}}\left[\widehat{\tilde{X}}'\tilde{\kappa}\widehat{\tilde{X}}\right]^{-1} \qquad (3.10)
\end{aligned}$$

with Ω_n the conventional diagonal matrix with elements taken as the squared residuals $(\widehat{\tilde{y}}_i - \widehat{\tilde{X}}_i'\hat{\beta}_n)^2$.

3.3 Semi-parametric efficiency

In order to achieve semi-parametric efficiency, Lavergne and Patilea (2013) extend the objective function (2.8) by considering a vanishing bandwidth associated with $k(.)$ and introducing a weighting matrix. Their efficient estimator is then obtained in two steps by using a consistent (first-step) estimator to consistently estimate the efficient weighting matrix. It does not seem however possible to design such a two-step procedure that would yield a semi-parametrically efficient R-SMD estimator. This is due to the Nadaraya-Watson estimator used in the Robinson's transformation to partial out the non-linear part of the model: the associated estimation error interacts with the efficient weighting matrix, resulting in a U-statistic with a non-zero mean that generates a bias. Such a bias involves first- and second-order derivatives of conditional expectations, and a bias-correction approach does not appear feasible either. Technical derivations can be found in the Supplementary Appendix.

Such an issue is not specific to our framework and appears more generally when a (first-step) nonparametric estimator is plugged into a moment function. The concern is then about *local robustness* with respect to the first step. To achieve semi-parametric efficiency, we suggest instead the following three-step approach:

1. compute the consistent R-SMD estimator $\hat{\beta}_n$ of β_0 defined in (3.9);

2. compute a nonparametric estimator for $g(.)$ using observations on y_i^* and Z_i with $y_i^* \equiv y_i - X_i'\hat{\beta}_n$;

3. compute the SMD estimator for β_0 using the locally robust moment function constructed by adding the influence function adjustment for the first-step estimator as shown in Chernozhukov et al. (2018)[7].

[7] We thank Chris Muris for bringing this work to our attention.

The detailed study of the properties of this three-step estimator is beyond the scope of our paper and is left for future research. Of course, efficient estimation of β_0 could be achieved by implementing the sieve minimum distance estimation procedure of Ai and Chen (2003) through a (sieve) approximation of the (unknown) $g(.)$ function in (2.1).

4 Simulation study

We investigate the small sample properties of our R-SMD estimator of β_0 in the following partially linear model,

$$y_i \quad = \quad X_i\beta_0 + g(Z_i) + e_i \quad \text{and} \quad X_i \;=\; f(W_i, Z_i) + v_i \tag{4.11}$$

where Z_i, X_i, and W_i are all univariate[8]. Our benchmark model labelled N-N model displays nonlinear first and second stages[9], respectively with nonlinear $f(.)$ and nonlinear $g(.)$ - e.g. cubic polynomial functions. The true unknown parameter of interest is β_0 and set at $\beta_0 = 2$ throughout. The instrument and covariate (W_i, Z_i) and the errors (e_i, v_i) are generated independently (for each i and from one another) according to two independent bivariate normal distributions with mean 0 and covariance matrix Σ_1 and Σ_2 respectively, with

$$\Sigma_1 = \begin{pmatrix} 1 & 5/9 \\ 5/9 & 1 \end{pmatrix} \qquad \text{and} \qquad \Sigma_2 = \begin{pmatrix} 1 & 4/9 \\ 4/9 & 1 \end{pmatrix}$$

This ensures that W_i and Z_i are exogenous, while X_i is endogenous. Our benchmark N-N model corresponds to the partially linear model (4.11) where

$$g(Z_i) \quad = \quad c_1 + \gamma_0 Z_i + \gamma_1 Z_i^2 \quad \text{and} \quad f(W_i, Z_i) \;=\; c_2 + \pi_0 W_i + \pi_1 W_i^2 + \alpha_0 Z_i + \alpha_1 Z_i^2 + \alpha_2 Z_i^3$$

When the parameters $\gamma_1, \pi_1, \alpha_1$, and α_2 are all set to 0, we have a L-L model. Otherwise, we have a N-N model when both $g(.)$ and $f(.)$ are non-linear, that is when γ_1 is non-zero and at least one non-zero parameter among π_1, α_1, and α_2.

We consider a sample of n i.i.d. observations on (y_i, X_i, Z_i, W_i). Our R-SMD estimator is computed using $\mu(.)$ chosen as the CDF of a standard Gaussian distribution[10]. The bandwidth of the Nadaraya–Watson estimation of the conditional expectations with respect

[8] We normalize X_i to ensure that its variance remains unchanged throughout the designs.

[9] In the supplementary appendix, a fully linear model with both first and second stages linear is also considered.

[10] Recall that the Fourier transform of a Gaussian function is also Gaussian, which is always greater than 0, symmetric, and available in closed-form. This implies that $\kappa_{j,l}$ (see (2.6)) is a real number; its expression is provided in the Supplementary Appendix. Other measures are also considered as a robustness check.

to Z_i is chosen according to the rule-of-thumb[11] (see Li and Racine (2007)), that is $\sigma_z n^{-0.2}$ with univariate Z_i. We compare the performance of our R-SMD estimator to three GMM-type estimators:

(i) the efficient GMM assumes that $g(.)$ is linear and relies either on one moment condition using W_i as the instrument, or two using W_i and W_i^2 as instruments;

(ii) the Sieves-GMM approximates $g(.)$ as a polynomial function in Z_i with its degree chosen by cross-validation; it relies either on one moment condition using W_i, or two using W_i and W_i^2 as instruments;

(iii) the R-GMM corresponds to the efficient GMM after a Robinson transformation; the same bandwidth is used for the Nadaraya-Watson estimation of the conditional expectation as for the R-SMD; it relies either on one moment condition using W_i, or two using W_i and W_i^2 as instruments.

The performance of these estimators is summarized by reporting the bias, standard error, and empirical rejection rates for a t-test of the null hypothesis $H_0 : \beta = \beta_0$ at 5% nominal level computed over 5,000 Monte-Carlo replications. The benchmark N-N model is generated as follows,

$$\begin{aligned} y_i &= 2X_i + 3Z_i - 3Z_i^2 + e_i \\ \text{with} \quad X_i^* &= W_i + 4W_i^2 + Z_i + 4Z_i^3 + v_i \quad \text{and} \quad X_i = 8\text{scale}(X_i^*) \end{aligned}$$

In Table 1.1, we report the Monte-Carlo bias and Monte-Carlo standard error (SE), as well as the average of the asymptotic SE assuming either homoskedasticity (Asympt. Homosk. SE) or heteroskedasticity (Asympt. Heterosk. SE), and the empirical rejection rates for a t-test at 5% nominal level of R-SMD and the 6 GMM-type estimators described above. We consider a sample of size $n = 200$ in Panel A and $n = 2,000$ in Panel B and $5,000$ Monte-Carlo replications. The bandwidth for the Nadaraya-Watson kernel estimators needed to compute R-SMD and R-GMM is set according to the rule-of-thumb, at 0.347 when $n = 200$ and at 0.219 when $n = 2,000$.

When the GMM-type estimators are computed using one moment (with instrument W) as is standard in practice, R-SMD outperforms them all and displays much smaller bias and standard errors both with small and large sample sizes. This is not surprising since important non-linearities are ignored by these GMM estimators. In Figures 1.1, we display the histograms of the Monte-Carlo distributions of these four estimators, as well as the distribution of their Monte-Carlo standard errors; for GMM, Sieves-GMM and R-GMM we consider either one moment using W or two moments using (W, W^2). In order to show the

[11] Alternative bandwidths were also considered as a robustness check: see the next subsection.

PANEL A: sample size $n = 200$

	R-SMD			R-GMM		Sieves-GMM		GMM	
Estimator	Gaussian $\mu(.)$	Cauchy $\mu(.)$	$sinc^2$ $\mu(.)$	W	(W, W^2)	W	(W, W^2)	W	(W, W^2)
Bias	0.011	0.009	0.003	0.095	-0.011	0.031	-0.007	0.404	-0.508
SE	0.036	0.038	0.041	3.031	0.033	3.129	0.067	16.812	0.444
Asympt. Homosk. SE	0.058	0.057	0.052	18.811	0.034	65.052	0.032	406.312	0.182
Asympt. Heterosk. SE	0.058	0.057	0.052	36.331	0.032	87.935	0.032	554.408	0.294
Rej. rate for Homosk. SE	0.004	0.006	0.013	0.007	0.046	0.006	0.072	0.010	0.924
Rej. rate for Heterosk. SE	0.004	0.006	0.016	0.000	0.072	0.002	0.087	0.007	0.763

PANEL B: sample size $n = 2,000$

Estimator	R-SMD			R-GMM		Sieves-GMM		GMM	
	Gaussian $\mu(.)$	Cauchy $\mu(.)$	$sinc^2$ $\mu(.)$	W	(W, W^2)	W	(W, W^2)	W	(W, W^2)
Bias	0.005	0.005	0.002	0.033	-0.006	0.052	0.000	0.006	-0.505
SE	0.010	0.011	0.012	0.069	0.009	0.167	0.009	0.832	0.093
Asympt. Homosk. SE	0.017	0.017	0.016	0.073	0.010	0.209	0.009	1.173	0.042
Asympt. Heterosk. SE	0.017	0.017	0.016	0.080	0.010	0.229	0.009	1.434	0.074
Rej. rate for Homosk. SE	0.004	0.005	0.010	0.038	0.068	0.022	0.048	0.021	1.000
Rej. rate for Heterosk. SE	0.004	0.005	0.010	0.008	0.074	0.009	0.049	0.011	0.999

Table 1.1: N-N model with $n = 200$ or $2,000$, and $M = 5,000$

We report the Monte-Carlo bias and Monte-Carlo standard error (SE), as well as the average of the asymptotic SE assuming either homoskedasticity or heteroskedasticity, and the empirical rejection rates of the null $H_0 : \beta = \beta_0$ using a 5% t-test for the R-SMD, R-GMM, Sieves-GMM and GMM estimators. R-SMD is computed with $\mu(.)$ chosen as the CDF of a standard Gaussian, Cauchy, or $sinc^2$ distribution. GMM, Sieves-GMM, and R-GMM are computed using either one moment with instrument W or two moments with instruments W and W^2.

magnitude and distribution of standard errors for different estimators, the histograms are drawn over the same range.

When considering two moments for the GMM-type estimators, their performances improve significantly. R-SMD still remains competitive with small bias, but its standard error is now slightly larger than R-GMM. Once again, this result is not surprising, and in line with our theoretical results. It is important to point out that the differences in standard deviations are quite small, both in small and large sample.

Finally, it is worth pointing out that R-SMD appears to be conservative with a rejection frequency under the null well below the nominal level equal to 5% - while other estimators appear slightly oversized. However, this under-rejection does not seem to be associated with poor power properties for R-SMD. In Figure 1.2, we display the power curves when testing the null $H_0 : \beta = \beta^*$ (and $\beta^* \neq \beta_0$) with t-tests using the different estimators at 5% nominal level. R-SMD remains competitive when compared to the most powerful procedures, R-GMM and Sieves-GMM using two moments.

- **Bandwidth selection**:

Next, we investigate the sensitivity of our results to the choice of the bandwidth. The results are displayed in Table 1.2 where we report the bias, standard error, and empirical rejection rates for a t-test at 5% nominal level as a function of the bandwidth that ranges from 0.087 to 0.347. The performance of our R-SMD estimator is remarkably stable: we only notice a small increase in the bias and a small decrease in the standard deviation as the bandwidth increases. We also report the results for a bandwidth selected by cross-validation[12] as well as the frequency with which each candidate bandwidth is selected: the rule-of-thumb bandwidth is selected most frequently, just under 80% of the time, and the associated performance results are quite similar to the previously reported ones.

Bandwidth	Frequency	Bias	SE	Asymp. Heterosk. SE	Rej. rate
0.08705506	0.159	-0.002	0.040	0.066	0.002
0.1519344	0.045	0.000	0.038	0.062	0.001
0.2168137	0.023	0.003	0.037	0.060	0.001
0.2816931	0.010	0.006	0.036	0.059	0.002
0.3465724	0.763	0.010	0.036	0.058	0.004
Cross-validation		0.010	0.037	0.060	0.003

Table 1.2: Simulation results for the N-N model with $n = 200$, $M = 5,000$

We report the Monte-Carlo bias and Monte-Carlo standard error (SE), the average of the asymptotic heteroskedasticity-robust SE, and the empirical rejection rates of the null hypothesis $H_0 : \beta = \beta_0$ using a 5% t-test for the R-SMD as a function of the bandwidth used in the Nadaraya-Watson estimate.

- **Selection of the measure $\mu(.)$ and associated $k(.)$:**

Last, we investigate the sensitivity of our results to the choice of the measure $\mu(.)$. As explained in section 3, the chosen measure $\mu(.)$ does not affect the consistency or the rate of convergence of the R-SMD estimator, but it may affect its asymptotic variance. The results are displayed in Table 1.1, where we report the performance of three R-SMD estimators obtained respectively with $\mu(.)$ chosen as the CDF of: (i) a standard Gaussian distribution and $k(.)$ Gaussian, (ii) a Cauchy distribution and $k(.)$ Laplace, and (iii) a $\texttt{sinc}^2$ distribution and $k(.)$ triangular.

Overall, the choice of the measure $\mu(.)$ does not seem to affect the performance of the R-SMD estimator much. In particular, the standard errors are quite similar across the three chosen measures, and even more so for the larger sample size $n = 2,000$.

- **Results for the L-L model:**

These results are reported in the Supplementary appendix. They show that the performance

[12] We select the bandwidth that produces the smallest Mean Squared Error (MSE). After splitting the sample into the training sample and the testing one, we compute the R-SMD estimate for a given bandwidth candidate, as well as the key parameters and the estimated $g(Z)$ function using a polynomial approximation. Then, we use these estimates to compute the MSE over the testing part of the sample.

of all considered estimators is quite similar with respect to bias, standard deviations and rejection rates.

5 Empirical application

To illustrate the applicability of our R-SMD estimator, we revisit and extend some of the empirical results in Dinkelman (2011) who estimates the impact of electrification on employment growth in South Africa. Because of measurement error and omitted variables concerns, the author adopts an instrumental variable empirical strategy based on the average community land gradient. Higher gradient raises the average cost of a household connection, making gradient an important factor in prioritizing areas for electrification. More specifically, the identification assumption is that conditional on baseline community characteristics, proximity to local economic centers and grid infrastructure, and district fixed effects, land gradient does not affect employment growth independently of being assigned an electrification project[13].

We focus on various outcome variables of interest that capture the effects of electrification on employment, household services, population growth, and skill composition of the labor force. Let y_i be the outcome variable of interest for community i, X_i the Eskom project variable which measures electrification, Z_i a vector of covariates[14], and $\tilde{W}_i$ the land gradient instrumental variable. We consider the partially linear IV model,

$$y_i = X_i\beta + g(Z_i) + e_i \qquad \text{where} \qquad E(e_i|Z_i, \tilde{W}_i) = 0 \tag{5.12}$$

We compare three estimation strategies:

(i) The empirical strategy in Dinkelman (2011) which relies on the standard IV method (labeled hereafter IV-L-L): it is based on a first stage linear in W and Z, and a linear second stage. In other words, model (5.12) is re-written as

$$\begin{aligned} y_i &= X_i\beta + Z_i'\gamma + e_i \qquad \text{where} \qquad E(e_i|Z_i, \tilde{W}_i) = 0 \\ X_i &= \pi W_i + Z_i'\delta + v_i \qquad \text{where} \qquad E(v_i|Z_i, \tilde{W}_i) = 0 \end{aligned}$$

[13] See sections 3 and 4 in Dinkelman (2011) for a detailed description of the data and the identification strategy.

[14] Covariates include household density, fraction of households living below a poverty line, distances to the grid, road, and town, fraction of adults that are white or Indian to proxy for local employers, fraction of men and women with a completed high school certificate, and two standard proxies for community poverty, the share of female-headed households and the female/male sex ratio (Guy Standing, John Sender, and Jeremy Weeks 1996). A set of ten district fixed effects are also included to ensure that all comparisons across project and non-project areas occur for areas in the same local labor markets. See Table 3 in Dinkelman (2011). For results obtained with an alternative set of control variables, see Supplementary Appendix.

(ii) The empirical strategy in Dieterle and Snell (2016) which relies on the IV method (labeled hereafter IV-Q-L): it is based on a first stage quadratic in W, linear in the controls, and as a linear second stage. In other words, model (5.12) is re-written as

$$\begin{array}{rcl} y_i &=& X_i\beta + Z_i'\gamma + e_i \qquad\qquad \text{where} \qquad E(e_i|\tilde{W}_i, Z_i) = 0 \\ X_i &=& \pi_1 W_i + \pi_2 W_i^2 + Z_i'\delta + v_i \qquad \text{where} \qquad E(v_i|\tilde{W}_i, \tilde{W}_i^2, Z_i) = 0 \end{array}$$

(iii) The R-SMD method proposed in this paper is based on model (5.12). Note that to handle the large number of covariates, we consider instead their first two principal components.[15]

In Table 1.3, we report the effect of electrification on employment (in Panel A) and on household energy sources and other household services (in Panel B); see also Tables 4, 5, and 8 in Dinkelman (2011)[16]. In Table 1.4, we report the effect of electrification on population growth, skill composition of labor force, and employment of incumbents; see also Table 10 in Dinkelman (2011). Overall, our R-SMD estimates tend to be smaller in magnitude and more precise (with smaller standard errors) than the ones obtained using IV-L-L and somewhat similar to the ones obtained using IV-Q-L with slightly larger standard errors. Specifically, we find no significant effect of electrification on employment (with or without in-migrants), but statistically and economically significant effect on all household energy sources and households services, on population growth (with and without in-migrants), as well as on the change in fraction of women (but not men) that have a completed high school education. Even if we cannot rule out that our R-SMD estimates are statistically different from those obtained using IV-L-L or IV-Q-L, there are some key differences that are worth highlighting.

First, in panel A of Table 1.3, we do not find any significant effect of electrification on employment rate[17]. Similar results are also obtained when considering the set of incumbents that excludes in-migrants (see panel B of Table 1.4). Our results are in line with those obtained using IV-Q-L while imposing a quadratic first stage, and suggest that there are important non-linearities in the model that need to be taken into account for reliable inference. That being said, it is important to distinguish electrification effect on female and

[15] The results for IV-L-L and IV-Q-L were obtained using the first two principal components. Also, the results for IV-L-L and IV-Q-L were compared between the two principal components and all controls. The results are very similar.

[16] To address concerns about *"overoptimistic inference with a possibly weak instrument"*, Dinkelman (2011) also reports in her Tables 4 and 5 identification-robust confidence intervals for the main Eskom project parameter estimate. We did not find much evidence of the presence of weak identification. Nonetheless, we report in the Supplementary Appendix corresponding identification-robust confidence intervals.

[17] The dependent variables represent the change in female (or male) employment rate between 1996 and 2001. Dinkelman (2011) found a significant effect on Female employment rate.

Panel A: Effects on employment

Outcome is Δ_t in	IV (L-L)	IV (Q-L)	R-SMD (NL-NL)
Female employment rate	0.090* (0.050) [-0.008,0.188]	0.053 (0.043) [-0.031,0.137]	0.045 (0.043) [-0.039,0.129]
Male employment rate	0.033 (0.062) [-0.089,0.155]	-0.013 (0.057) [-0.125,0.099]	0.061 (0.059) [-0.055,0.177]

Panel B: Effects on Household energy sources & other household services

Outcome is Δ_t in	IV (L-L)	IV (Q-L)	R-SMD (NL-NL)
Lighting with electricity	0.642*** (0.176) [0.297,0.987]	0.368*** (0.114) [0.145,0.591]	0.359** (0.145) [0.075,0.643]
Cooking with wood	-0.282** (0.125) [-0.527,-0.037]	-0.221** (0.097) [-0.411,-0.031]	-0.202* (0.123) [-0.441,0.037]
Cooking with electricity	0.239*** (0.080) [0.082,0.396]	0.141*** (0.054) [0.035,0.247]	0.152** (0.074) [0.007,0.297]
Water nearby	-0.372* (0.197) [-0.758,0.014]	-0.363** (0.167) [-0.690,-0.036]	-0.626** (0.246) [-1.108,-0.144]
Flush toilet	0.067 (0.055) [-0.041,0.175]	0.069 (0.052) [-0.033,0.171]	0.104* (0.060) [-0.014,0.222]

Table 1.3: Impact of electrification on Employment (Panel A) and on Household energy sources & other household services (Panel B)

Note: *** Significant at 1%, ** at 5%, * at 10%. Each cell in the table presents estimates of the Eskom project variable coefficient, robust standard error, and 95% confidence interval from an IV regression of the dependent variable on the Eskom project indicator and control variables (that include baseline controls and district fixed effects; see Table 3 in Dinkelman (2011)). In Panel A, the dependent variable is the change in female (or male) employment rate between 1996 and 2001; in Panel B, the outcome variables measure the change in fraction of households using different energy sources or with access to basic services. Each regression contains $N = 1,816$ except for change in fraction of households using wood which contains $N = 1,807$ due to missing data on this variable.

male employment rate. For female employment rate, our estimate and standard error are very similar to the IV-Q-L ones, suggesting that non-linearities are mainly present in the first stage and adequately captured by a quadratic function. However, the picture is quite different when considering male employment rate where our estimate, while insignificant, has the expected positive sign unlike the IV-Q-L's; such a positive estimate is also obtained on the sample that excludes in-migrants, both using R-SMD and IV-Q-L. These differences suggest that a quadratic first stage is inappropriate. Additional estimates[18] obtained using R-GMM (with up to 3 powers of W as instruments) are also positive and insignificant, except when W is used alone. Even if it remains unclear whether the negative estimate obtained using IV-Q-L is driven by *missed nonlinearities* in the first or second stage, our results highlight the convenience of our R-SMD estimation strategy and suggest that it delivers reliable and precise estimates regardless[19].

Second, in panel B of Table 1.3, we find significant effect of electrification on all five measures of household energy sources and other household services[20]. Specifically, for the first three household energy sources, we report important shifts towards using electricity for home production (increase of 35.9%) and for cooking (increase of 15.2%), and a drop of 20.2% in relying on wood for cooking. Our estimates are smaller in magnitude than those obtained by Dinkelman (2011), but they are still statistically significant and economically meaningful. In addition, and unlike Dinkelman (2011), we find evidence that electrified regions also experience differential changes in the last two (basic) household services, namely access to piped water close to home and access to a flush toilet at home. Our estimates are larger in magnitude than both IV-L-L and IV-Q-L estimates and statistically significant. These differences suggest once again important nonlinearities that are not appropriately accounted for by either IV-L-L or IV-Q-L.

Third, in Table 1.4, we find significant effect of electrification on population growth (with and without in-migrants). The estimates obtained by all three methods may appear surprisingly large: even after accounting for in-migrants, electrified areas are found to have significantly higher population growth[21] (of the order of 300 percent!). We also find significant effect of electrification on the change in fraction of women that have a completed high school education, but not in fraction of men. Our results are in-line with those obtained using IV-Q-L, but not using IV-L-L where significant effects on both women and men are reported. However, the significance of the IV-L-L estimate reported for men is fragile and disappear when considering a slightly different set of controls whereas all other results remain stable (see Supplementary Appendix). In addition, it is important to highlight that

[18] See Supplementary Appendix.

[19] Differences between R-SMD and TSLS estimates could also result from using the wrong functional form of the structural equation or heterogeneous effects. As a result, our interpretations should be taken with a grain of salt without further investigation.

[20] The outcome variables measure the change in the fraction of households using different energy sources or with access to basic services.

[21] As explained in Dinkelman (2011) on p3103, "*in small communities, numerically small increases in population can translate into large percentage changes. The average number of females (males) in these communities in 1996 is 356 (274). This rises to 446 (319) by 2001. Just considering the raw changes in number of adults over time, electrified areas grow at about 6% per year while non-electrified areas grow at about 3%.*"

Panel A								
Δ_t log population			Δ_t Females with High School			Δ_t Males with High School		
IV (L-L)	IV (Q-L)	R-SMD (NL-NL)	IV (L-L)	IV (Q-L)	R-SMD (NL-NL)	IV (L-L)	IV (Q-L)	R-SMD (NL-NL)
3.820***	3.101***	2.949***	0.128***	0.102***	0.195**	0.076*	0.047	0.089
(0.987)	(0.710)	(0.841)	(0.048)	(0.039)	(0.082)	(0.043)	(0.037)	(0.057)
[0.885, 5.755]	[1.710,4.493]	[1.301,4.597]	[0.034,0.222]	[0.026,0.178]	[0.034,0.356]	[-0.008,0.160]	[-0.026,0.120]	[-0.023,0.201]
Panel B								
Δ_t log non in-migrant population			Δ_t Females empl. excl. in-migrants			Δ_t Males empl. excl. in-migrants		
IV (L-L)	IV (Q-L)	R-SMD (NL-NL)	IV (L-L)	IV (Q-L)	R-SMD (NL-NL)	IV (L-L)	IV (Q-L)	R-SMD (NL-NL)
4.275***	3.348***	3.215***	0.113**	0.074*	0.046	0.087	0.026	0.093
(1.092)	(0.753)	(0.908)	(0.051)	(0.042)	(0.041)	(0.063)	(0.054)	(0.060)
[2.135, 6.415]	[1.872,4.824]	[1.435,4.995]	[0.013,0.213]	[-0.008,0.156]	[-0.034,0.126]	[-0.036,0.210]	[-0.080,0.132]	[-0.025,0.211]

Table 1.4: Impact of electrification on Population growth, skill composition of labor force and employment of incumbents

Note: *** Significant at 1%, ** at 5%, * at 10%. Each cell in the table presents estimates of the Eskom project variable coefficient, robust standard error, and 95% confidence interval from an IV regression of the dependent variable on the Eskom project indicator and control variables (that include baseline controls and district fixed effects; see Table 3 in Dinkelman (2011)). Dependent variable in panel A, column 1, is change in log African population; in columns 2-3 it is the change in fraction of women or men that have a completed high school education. Dependent variable in panel B, column 1, is the change in log African non–in-migrant population where in-migrants have been subtracted from the total number of adults in the community in each year. In columns 2-3 of panel B, the outcomes are change in female and male employment rates where the employment variables exclude the number of in-migrants to each community in each year. Each regression contains $N = 1,816$.

in both cases, R-SMD estimates are actually larger in magnitude than the ones obtained using IV-L-L and IV-Q-L. Larger estimates are also reported using R-GMM (with up to 3 powers of W as instruments), suggesting the presence of nonlinearities in the model that need to be taken into account for reliable inference.

Overall, our empirical results highlight the importance of using a flexible and user-friendly estimation strategy such as R-SMD to deliver reliable inference without having to rely on a potentially problematic and misleading parametrization of the first stage, or the second stage with respect to the controls.

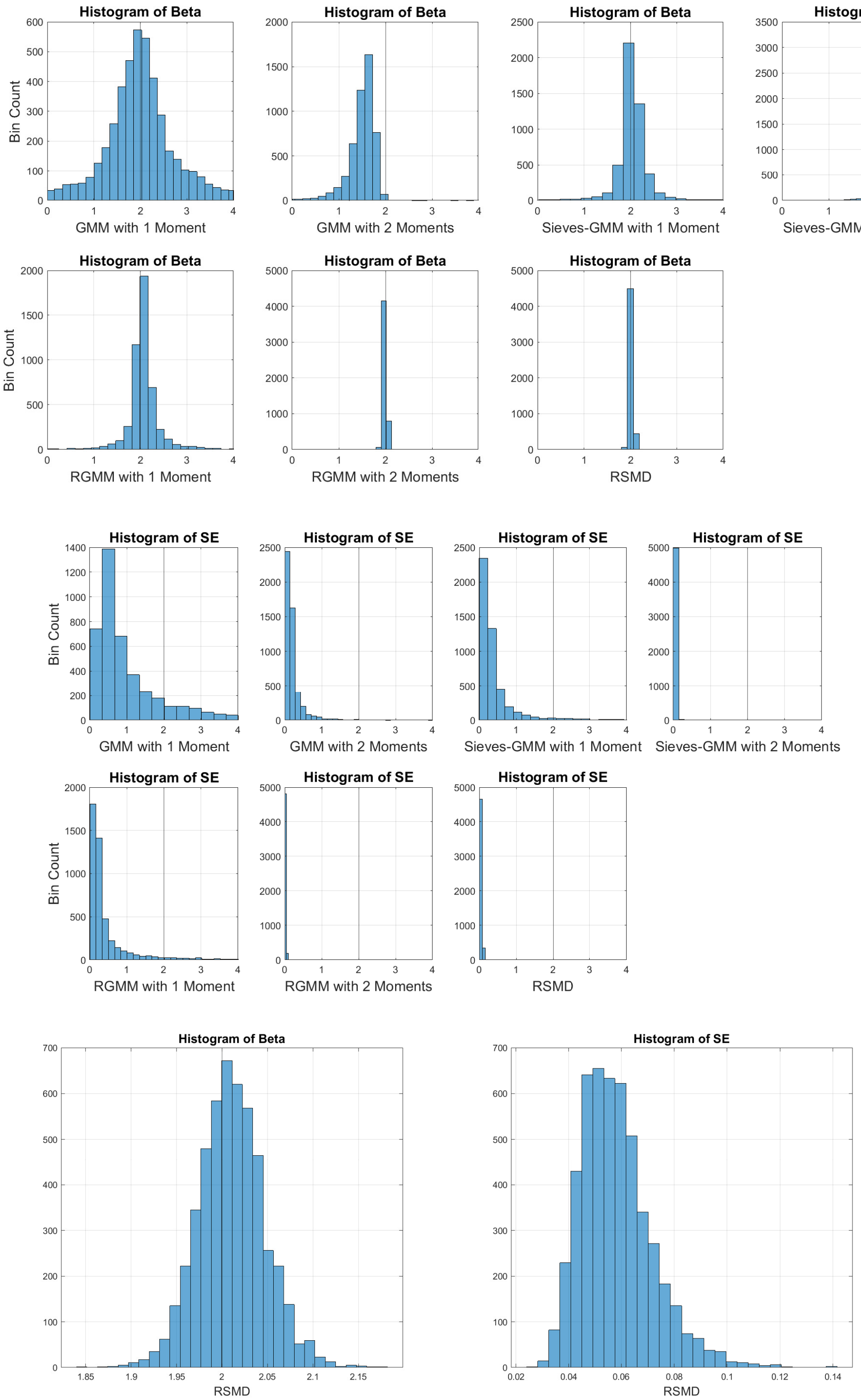

Figure 1.1: Simulation Results for the N-N model using $n = 200$ and $M = 5,000$

We display the Monte-Carlo distribution of the following estimates (top 2 rows) and of their standard deviations (rows 3 and 4): from left to right, top row, GMM with 1 and 2 moments and Sieves-GMM with 1 and 2 moments; bottom row, R-GMM with 1 and 2 moments and R-SMD. On the bottom row, we display the Monte-Carlo distribution of the R-SMD estimates and of their standard deviations.

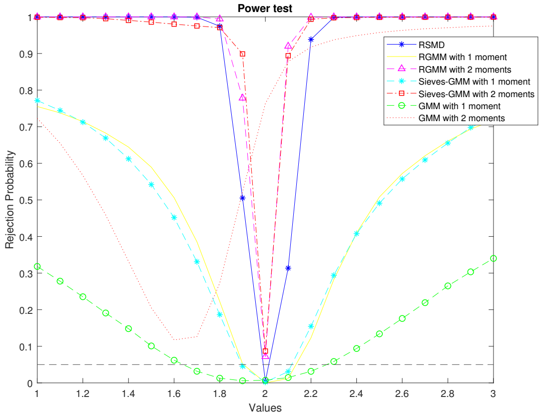

Figure 1.2: Power curves for the N-N model with $n = 200$, $M = 5,000$

GMM, Sieves-GMM and R-GMM with 1 and 2 moments, and R-SMD.

Chapter 2

Estimation of Heterogeneous Treatment Effects Using a Conditional Moment Based Approach

1 Introduction

Heterogeneous treatment effects have been a focus in literature since Imbens and Angrist (1994). Heterogeneous treatment effects models are those where the treatment effects of policies, programs, or other variables on some outcome of interest may vary across individuals with different characteristics. To measure heterogeneous treatment effects, many papers add interaction terms between the treatment variable and some covariates to the standard linear model (see Wooldridge (2010) and Imbens and Rubin (2015)). The basic linear model becomes linear in the treatment, the interaction term, and the covariates. The parameters of interest are the parameter for the treatment and that for the interaction term.

Unfortunately, the true model is unknown. If the model is nonlinear in the covariates, OLS, 2SLS, or GMM with a linear form will suffer from model misspecification, and estimates will not be consistent and normally distributed. With the correct model form, the OLS, 2SLS, or GMM will work. However, if the nonlinear part of the model is unknown, we will need to use other methods to estimate the parameters. There are at least three challenges associated with the estimation of both parameters in the model with a parametric part and unknown nonlinear part, i.e., the partially linear model (PLM).

One challenge happens when the treatment variable is endogenous, we need to use valid instruments to estimate parameters for the treatment and for the interaction term. The usual approach would be to partial out the unknown and nonlinear part of the model via a Robinson (Robinson (1988)) transformation and then use the GMM estimator with one instrument and the interaction term between the instrument and the covariate to estimate both parameters. When we only have one valid instrument and the covariate inside the interaction term has little variation or is not very informative, e.g., rare program participation, the GMM estimator will not be consistent and its variance will have a large magnitude. This is similar to a weak identification problem. Additionally, the selection of instruments will affect the interpretation of the treatment (see Dieterle and Snell (2016) for related discussions).

A second challenge is that traditional estimation methods do not work when the number of covariates is possibly larger than the number of observations. This challenge exists if there are a large number of covariates in the structural model or if many covariates must be included to ensure that the instrument is valid.

The third challenge is that if the covariates enter the model in an unknown (and possibly nonlinear) way, estimation methods assuming linearity will suffer from specification bias.

Our new estimator solves the first challenge by using only the valid instrument (e.g., the random assignment) to estimate both parameters. Our method does not use the interaction term between the instrument and the covariate to identify model parameters. We do this by first applying the Robinson transformation to remove the nonlinear and unknown part, and then transforming the conditional moment restriction: $E(\epsilon_i|Z_i) = 0$ where ϵ_i is the error term inside the model and Z_i is the instrument. Usually, GMM uses the unconditional moments generated by $E(\epsilon_i|Z_i) = 0$ to construct objective functions. But, with the i.i.d assumption we can also generate objective function as $E(\epsilon_i|Z_i)E(\epsilon_j|Z_j)$. As written, this type of objective function is hard to use. To make it tractable, we apply both the i.i.d. assumption and the Fourier transform to obtain a new objective function related with $E(\epsilon_i|Z_i)E(\epsilon_j|Z_j)$. This is tractable. In technical terms, the Fourier transform and the new objective function are based on the Smooth Minimum Distance (SMD) approach (Lavergne and Patilea (2013)). The SMD-based approach exploits all the information of the conditional mean independence restriction without having to specify the first stage, or to select a finite number of unconditional moments (e.g., using transformations of Z_i).

To tackle the last two challenges, our new estimator uses regularized Machine Learning (i.e., the Lasso method) to estimate the nuisance parameters generated by Robinson transformation (Robinson (1988)) and a Neyman-Orthogonalized (Chernozhukov et al. (2018)) First-Order Condition (FOC). With the Robinson transformation, we do not need to specify how these covariates enter the model; it partials out the unknown (and nonlinear) part of the partially linear model, but it also introduces nuisance parameters which need to be estimated. The Neyman-Orthogonalized FOC reduces the effect of the bias associated with the estimation of these nuisance parameters. When there are many covariates, assuming sparsity and using regularized machine learning methods, such as the Lasso method, allows us to choose a relatively small number of covariates for the estimation of nuisance parameters. Thus, our new estimator is the D-RSMD estimator (Debiased Robinson-SMD).

To apply the new estimator we only need one tuning parameter. If we are using the Lasso method, the tuning parameter is inside the penalty term for the estimation of the nuisance parameter. In practice, choosing Lasso tuning parameters is well-understood (see van de Geer (2016)) and we use cross-validation (Chernozhukov et al. (2018)). When there are few covariates (and the model is sparse by design), the Nadaraya–Watson estimator is an alternative method to estimate the nuisance parameter. This estimation procedure also needs one tuning parameter, that is, the bandwidth. The bandwidth can be chosen by the rule of thumb or cross-validation in practice. We show that our D-RSMD estimator is consistent and $\sqrt{n}$ asymptotically normal under mild regularity conditions. Monte-Carlo simulations show that the D-RSMD estimator outperforms the R-SMD and GMM-type estimators in terms of much smaller bias and lower standard errors.

To illustrate the value of our method, we estimate both parameters inside the heterogeneous treatment effects to extend the estimation results in Card (1993). We use the Card application to show the differences of estimators using various sets of instruments. It shows that our new method generates more precise estimates in comparison to GMM.

Our work is an extension of Antoine and Sun (2021) who propose the R-SMD estimator which combines a Robinson transformation and a smooth minimum distance estimator in a partially linear model[1]. However, when we use the Nadaraya-Watson estimator to estimate the conditional means, the dimension of covariates has to be less than four (see chapter 7 of Li and Racine (2007)). To extend the R-SMD estimator to the big data setting and relax the restriction on the dimensions of covariates, our work applies machine learning methods, such as the Lasso. To reduce the effect of bias introduced by Lasso, our work employs the Neyman-orthogonal first-order condition. The method also extends the Neyman-orthogonal estimator (Chernozhukov et al. (2018)) to the U-statistic setting. Our new estimator contributes to the literature on the partially linear model and conditional moment restriction models in the same way as the R-SMD estimator does. Both the partially linear model and conditional moment restriction model involve the problem of choosing the model form. For a partially linear model, some researchers use sieves for the nonlinear part. For conditional moment restriction models, some generate a finite number of unconditional moments. In comparison, our method does not need to choose the number of polynomials for the sieve method or the number of unconditional moments. Our method only needs to choose one tuning parameter.

The paper is organized as follows. Section 2 introduces our framework, motivation, and our estimator. Section 3 states the large sample properties for our estimator. Section 4 presents the simulation results for finite samples with and without a large number of covariates. Section 5 uses the D-RSMD estimator to estimate the effects of education on earnings. The additional application results and proofs are in the Appendix and Supplementary Appendix.

[1] See Robinson (1988) and Li and Racine (2007) for combining Robinson transformation and 2SLS estimators.

2 Framework and Motivation

Our framework is based on a partially linear model with a treatment variable W_i which is binary. The heterogeneous treatment effects enter the model through the linear term on W_i and some interaction terms between W_i and X_i, a vector of P covariates with $X_i = [X_{i1}', X_{i2}']'$. The other part of the model is an unknown function of X_i.

$$y_i = \theta_{w0} W_i + W_i \cdot X_{i1}' \theta_{wx0} + f_{0,1}(X_i) + \epsilon_i \tag{2.1}$$

The treatment is not always exogenous. For example, in the Oregon Medicaid health experiment, the treatment is the enrollment in Medicaid. Enrollment in Medicaid is endogenous because it is based on the choices of the lottery winners. Only a fraction of lottery winners decided to enroll in Medicaid.

$W_i \cdot X_{i1}$ is the interaction term. The formal definition for $W_i \cdot X_{i1}' \theta_{wx0}$ is

$$\sum_{k=1}^{p_1} W_i X_{i1,k} \times \theta_{wx0,k}$$

where $X_{i1,k}$ is a k-th element inside the vector X_{i1}. The treatment part of interest is $W_i + W_i \cdot X_{i1}$. The parameters that measure heterogeneous treatment effects, i.e. θ_{w0} and θ_{wx0}, are our key parameters. The control vector X_i in our setting is exogenous and appears in an unknown function $f_{0,1}(X_i)$. $f_{0,1}(.)$ is a nuisance parameter that we are not interested in.

X_{i1} is a subvector of X_i of dimension p_1 with $p_1 \leq P$. We can have $X_{i1} = X_i$ if the dimension of X_{i2}, p_2, is zero. That is, $W_i \cdot X_{i1}$ can be the interaction term using only a small set of covariates, while $f_{0,1}(X_i)$ is the unknown function of the whole set of exogenous controls. The dimension of X_i is not restricted and can be either high or low-dimensional. However, we restrict the dimension of X_{i1}. And we also maintain the exogeneity of all the controls X_i. The outcome variable y_i is not necessarily binary. For a binary outcome variable, a large number of covariates, and similar restricted dimension setting on estimating heterogeneous treatment effects, see Nekipelov et al. (2018).

The $f_{0,1}(X_i)$ is unknown allowing X_i to enter the model in a flexible way. To avoid estimating the nuisance parameter $f_{0,1}(X_i)$, a Robinson transformation is introduced to eliminate the unknown $f_{0,1}(X_i)$. The Robinson transformation consists in subtracting the conditional expectation of y_i with respect to the controls X_i. After such a Robinson transformation, $f_{0,1}(X_i)$ will disappear. In this case, we do not need to assume the form of $f_{0,1}(X_i)$. It can be sparse and may be nonlinear. Under Robinson transformation and the exogenous assumption on X_i, Equation 2.1 becomes

$$y_i - E(y_i|X_i) = (P_i - E(P_i|X_i))' \theta_0 + \epsilon_i$$

where $P_i = [W_i, W_i \cdot X_{i1}']'$ and $\theta_0 = [\theta_{w0}, \theta_{wx0}']'$.

Denote

$$\epsilon_j(\theta, g_0) = \tilde{y}_j - \tilde{P}_j\theta \tag{2.2}$$

where $\tilde{y}_j \equiv y_j - E(y_j|X_j)$ and $\tilde{P}_j \equiv P_j - E(P_j|X_j)$. g_0 stands for all of the unknown real values of nuisance parameters. g_0 is a vector of $g_{0,y}(X_j)$ (or $E(y_j|X_j)$) and $g_{0,P}(X_j)$ (or $E(P_j|X_j)$) at this stage.

As in the previous discussion, W_i and the interaction term between W_i and X_{i1} are all endogenous. Following the classic endogenous variable estimation, a vector of instruments, including the random assignment Z_i, is introduced. For instance, in the Oregon Medicaid health experiment, the instrument is the lottery outcome[2]. We maintain the conditional mean independence assumption for the random assignment. With valid instrument and exogenous control variables, the conditional moment restriction is

$$E(\epsilon_i|X_i, Z_i) = 0 \ a.s. \tag{2.3}$$

In our setting, we look at the single treatment case where the controls may be correlated with the instruments. With traditional estimators such as GMM or 2SLS, $1+p_1$ instruments are needed to identify (and estimate) the slope parameters associated with W_i and the interaction terms $W_i \cdot X_{i1}$. GMM will use Z_i (the random assignment) and the interaction terms $Z_i \cdot X_{i1}$ to estimate the key parameters.

However, if there is little variation in X_{i1}, the $Z_i \cdot X_{i1}$ is less informative, or if we use a second instrument that is (highly) correlated with the valid Z_i, the GMM method will fail to provide reliable estimates: intuitively, it is similar to a weak instrument problem. With only one valid instrument, GMM encounters the problem of how to generate new moments when we need to estimate more than one parameter. With only one valid instrument and two parameters to estimate, there is under-identification, and GMM cannot be implemented. We will show that our estimation strategy, which only relies on using one valid instrument (e.g. the random assignment Z_i), delivers reliable inference on both parameters.

Further, when it comes to the interpretation of the heterogeneous treatment effects, when using two or more instruments, additional assumptions are needed because the traditional monotonicity assumption (as in Imbens and Angrist (1994)) is only for one instrument. For our setting, in the simplest case where X_{i1} is one dimensional (with values -1 or 1), if we use instruments Z_i (with values 0 or 1) and $Z_i \cdot X_{i1}$, then there are two types of compliers. The first kind of compliers are the individuals who will accept the treatment if and only if Z_i is one regardless of the values of X_{i1}. The second kind of compliers are the individuals who will accept the treatment if and only if Z_i is one and X_{i1} is positive.

[2] In the experiment, the lottery winners are allowed to enroll in the Medicaid program. Not every lottery winner enrolled. The treatment variable is the actual enrollment status.

Hence, if we want to use LATE interpretation as in Imbens and Angrist (1994) then we cannot include the interaction term to measure the heterogeneous treatment effects. If we need to estimate the interaction term, then we need to use the interaction term between the instrument and covariates as extra instruments. As a consequence, the types of compliers will affect the interpretation of LATE. This interpretation problem is discussed further after our formal Assumption $4(v)$ is introduced.

This dilemma can be solved by employing the conditional moment restriction directly. The conditional moment restriction contains all of the information no matter the number of instruments inside. Hence, the conditional moment based approach enables us to use one instrument Z_i (e.g. the random assignment) only to identify and estimate slope parameters associated with both W_i and $W_i \cdot X_{i1}$. The Bierens type estimator is a conditional moment based approach which transforms the conditional moment into an infinite number of unconditional moments using complex exponential functions. See Bierens (1982), Antoine and Lavergne (2014), and Antoine and Sun (2021).

$$E[\epsilon_j(\theta_0, g_0)e^{it'Z_j}] = 0 \; \forall t \in \mathbb{R}^{q_z} \qquad \iff \qquad E(\epsilon_j(\theta_0, g_0)|Z_j) = 0 \; a.s. \tag{2.4}$$

The population objective function defined in Equation (2.5) below is based on the previous equation. Inside the objective function, $\mu(t)$ is a strictly positive measure on the vector t. Z_i stands for the vector of instruments. [3]

$$M_\infty(\theta, g) = \int_{\mathbb{R}^{q_z}} |E[\epsilon_j(\theta, g)e^{it'Z_j}]|^2 d\mu(t) \tag{2.5}$$

The objective function involves the norm of a complex function. To estimate θ_0 we need to find the derivative of the norm of the complex function, which is difficult to compute. Under the independent assumption for the population, the objective function has an alternative expression.

$\forall j \neq l$,

$$M_\infty(\theta, g) = E[\epsilon_j(\theta, g)\epsilon_l(\theta, g)\kappa_{j,l}] \;\text{ with }\; \kappa_{j,l} = \int_{\mathbb{R}^{q_z}} e^{it'(Z_j - Z_l)} d\mu(t) \tag{2.6}$$

The objective function defined in Equation (2.6) is a function of the parameters θ we are interested in and the nuisance parameters g as in the Equation (2.2) after the Robinson transformation. With Regularity Assumptions provided in the following, θ_0 is the unique minimizer of the objective function $M_\infty(\theta, g_0)$ where g_0 is a vector of $g_{0,y}(X_j)$ (or $E(y_j|X_j)$) and $g_{0,P}(X_j)$ (or $E(P_j|X_j)$) at this stage.

[3] Z_i denotes the general vector of instruments. If we only use one instrument, Z_i will be the random assignment. If we use a vector of instruments, Z_i will be the random assignment and the interaction term between random assignment and covariates.

Assumption 4. *(Regularity assumptions)*
(i) $E(\epsilon_i|X_i, Z_i) = 0$.
(ii) $E(\tilde{P}_i|Z_i) \neq 0$ *a.s. (with probability 1) with* $\tilde{P}_i = P_i - E(P_i|X_i)$.
(iii) $E(\tilde{P}_i\tilde{P}_i')$ *is nonsingular.*
(iv) Let $f_Z(.)$ *denote the density function of* Z_j. *We assume that* $E(\tilde{P}_j|Z_j = .)f_Z(.)$ *is* L_q *for some* $1 \leq q \leq 2$.
(v) for all possible z_1, z_2, *either* $E(W_i|Z_i = z_1) \geq E(W_i|Z_i = z_2)$ *for all* i, *or* $E(W_i|Z_i = z_1) \leq E(W_i|Z_i = z_2)$ *for all* i.
(vi) (y_l, W_l, X_l, Z_l) *is an independent and identical copy of* (y_j, W_j, X_j, Z_j).
(vii) Let μ *be a given strictly positive measure on* $\mathbb{R}^{q_z}$. *Let* $k(.)$ *be the Fourier transform induced by* μ, $k(Z_j - Z_l) = \int_{\mathbb{R}^{q_z}} e^{it'(Z_j - Z_l)}d\mu(t)$. *We assume that* $k(.)$ *is a symmetric bounded density function on* $\mathbb{R}^{q_z}$ *and that its Fourier transform is strictly positive.*

Assumption 4(i) is the exogeneity assumption for the controls and instruments. In the potential outcomes setting, Assumption 4(i) is also the Random Assignment Assumption. The model form implies the Exclusion Restriction Assumption as in Imbens and Rubin (2015), that is, the value of the instrument does not affect the potential outcomes directly. Assumption 4(ii) is the relevant instrument assumption. Assumption 4(iii) and 4(iv) guarantee the identification of the parameters that we are interested in. Robinson (1988) also imposes the same assumption as Assumption 4(iii). Assumption 4(vii) is for the measure $\mu(.)$. These conditions in Assumption 4(vii) are not very restrictive. There are many different available measures. We use the CDF of Gaussian distribution in simulations and applications.

The Monotonicity Assumption (or Assumption 4(v)) needs more investigation. Assumption 4(v) is a multivariate extension of the Monotonicity Assumption from Imbens and Angrist (1994). It is a condition that assumes all individuals make the same choice if they are given the same options. That is, even if there are two types of compliers in reality, Assumption 4(v) assumes that only one type exists. Recall that when individuals accept the treatment based on the value of the covariate inside the interaction term, they are different compliers. If we only use one instrument, Assumption 4(v) will simply be the Monotonicity Assumption. The interpretation of the parameters θ_{w0} and θ_{wx0} are associated with the average treatment effect of the compliers who make their choices based on the result of the random assignment. If Z_i is a vector of instruments, Assumption 4(v) still ensures that there is only one type of compliers. The interpretation of the parameters θ_{w0} and θ_{wx0} is connected with the average treatment effect of that kind of compliers. We will show that our estimator can be reliably implemented with only one instrument. In our simulations and applications, we will provide results using one and multiple instruments for comparison.

Under Assumption 4, θ_0 is the unique minimizer of

$$M_\infty(\theta, g) = 0$$

when $g = g_0$, because $E(\epsilon_i|Z_i) = 0$ with probability 1. Without replacing the nuisance parameter g with its true value g_0 the First Order Condition of $M_\infty(\theta, g)$ is

$$E((P_j - g_P(X_j))\,[y_l - g_y(X_l) - (P_l - g_P(X_l))'\theta]\kappa_{j,l}) = 0 \tag{2.7}$$

When $g = g_0$, $g_P(X_j)$ becomes $g_{0,P}(X_j)$ or $E(P_j|X_j)$ and $g_y(X_l)$ is $g_{0,y}(X_l)$ or $E(y_l|X_l)$, the FOC becomes:

$$E[\tilde{P}_j(\tilde{y}_l - \tilde{P}_l'\theta)\kappa_{j,l}] = 0 \tag{2.8}$$

θ_0 is identified under a strong identification assumption and Assumption 4. We leave the formal discussion to Proposition 3. The FOC defined in Equation (2.8) provides an explicit form for θ_0. The sample analog of the explicit form for θ_0 under Equation (2.8) is a direct extension from Antoine and Sun (2021) by allowing interaction term inside the parametric part of the model and the expression for θ_0 is provided in the Appendix Section 3. Antoine and Sun (2021) provide the estimator for the parameter in front of the treatment. It is a special case of Lavergne and Patilea (2013) when the bandwidth inside $\kappa_{j,l}$ is fixed. Lavergne and Patilea (2013) show that such an estimator is consistent and asymptotically normal. However, since the nuisance parameter g_0 is unknown, the estimator based on the sample analogue of Equation (2.8) is infeasible and denoted as $\tilde{\theta}_{n,u}$. To obtain the feasible estimator, we need to estimate the nuisance parameters in the first step. Antoine and Sun (2021) replace their nuisance parameters with Nadaraya-Watson estimators, and show that with proper assumption on the bandwidth, the infeasible and feasible estimators share the same asymptotic properties. To satisfy the assumption on bandwidth, the Nadaraya-Watson estimator imposes one constraint on the number of covariates. We will discuss this constraint in the later sections about feasible estimators. Using Nadaraya-Watson estimators or other non-parametric estimators, with regularized conditions, such as the Lasso method, the bias introduced in estimating the nuisance parameters may cause bias in the second stage where we estimate the key parameters.

We propose a new FOC to estimate the θ_0, which extends the Neyman-orthogonal estimator (Chernozhukov et al. (2018)) to the U-statistic setting. Our original FOC defined in Equation (2.7) is not orthogonal to the nuisance parameter, so the bias in the first step will affect the estimate of the key parameter. This is shown in the Appendix. The Neyman-orthogonal method is constructing a new FOC which is orthogonal to the nuisance parameters introduced in the first step. The new FOC has all partial derivatives with respect to the nuisance parameters equal to zero, in this way it is orthogonal to the bias of the nuisance parameter (or function). The definition of the partial derivative with respect to a function is provided in Chernozhukov et al. (2018). A similar method is used in Nekipelov et al. (2018).

After Neyman-orthogonalization, the new FOC for θ_0 becomes insensitive to the bias of the estimator for nuisance parameters. It is sensitive to the square of the bias. As long as the bias is $o_p(n^{-1/4})$, we still have a $\sqrt{n}$-asymptotically normally distributed estimator for the key parameter. This order for the bias is not an issue, because there are still many estimation methods to choose from, for instance, the Lasso, Sieves, Random forest, and so on.

The new FOC for θ_0 corresponding to Equation (2.7) is in the following.

$$\begin{aligned} E[\Psi(D;\theta,g_0)] &\equiv E\left[\left(P_j - g_{0,P}(X_j) - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)[y_l - g_{0,y}(X_l) - (P_l - g_{0,P}(X_l))'\theta]\kappa_{j,l}\right] \\ &= 0 \end{aligned} \tag{2.9}$$

with $g_{0,\tilde{P}_m}(X_l) \equiv E[(P_m - g_{0,P}(X_m))\kappa_{m,l}|X_l]$ and $g_{0,\kappa_{m,l}}(X_l) \equiv E[\kappa_{m,l}|X_l]$.

$g_{0,\tilde{P}_m}(X_l)$ and $g_{0,\kappa_{m,l}}(X_l)$ are two additional parameters inside the nuisance parameter vector. At this stage, there are four nuisance parameters inside the vector. With extra two nuisance parameters, the FOC defined in Equation (2.9) has a partial derivative with respect to all nuisance parameters equal to zero. This is shown in the Appendix.

Equation (2.9) gives us the identification of the true key parameter and the forms of the infeasible and feasible estimators.

Proposition 3. *(Identification of θ_0 using the orthogonalized FOC)*
Under Assumption 4 and FOC defined in Equation (2.9)

$$\theta_0^* = E\left[\kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)\tilde{P}_l'\right]^{-1} E\left[\kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)\tilde{y}_l\right]$$

If we plug Equation (2.2) into the formula for θ_0^* with error term ϵ_l, we will have $\theta_0^* = \theta_0$. The sample analog of θ_0^* under Equation (2.9) (in Proposition 3) delivers an estimator for θ_0. Because we have two distinct individuals inside the expectation (e.g., j and l), we need to replace the expectation by the average of a double summation to obtain the infeasible estimator under Equation (2.9). The closed-form expression for the infeasible estimator, $\tilde{\theta}_{n,o}$, is:

$$\begin{aligned} \tilde{\theta}_{n,o} = & \left[\tfrac{1}{n(n-1)}\textstyle\sum_{j=1}^n \sum_{l\neq j}^n \kappa_{j,l}\left(\tilde{P}_j - \tfrac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)\tilde{P}_l'\right]^{-1} \\ & \left[\tfrac{1}{n(n-1)}\textstyle\sum_{j=1}^n \sum_{l\neq j}^n \kappa_{j,l}\left(\tilde{P}_j - \tfrac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)\tilde{y}_l\right] \end{aligned}$$

To distinguish the infeasible estimators under two different FOCs, the infeasible estimator under orthogonal FOC (Equation (2.9)) is denoted as $\tilde{\theta}_{n,o}$. Recall that the infeasible estimator under Equation (2.8) (from Antoine and Sun (2021)) is $\tilde{\theta}_{n,u}$ and the expression is shown in the Appendix Section 3.

3 Large Sample Theory

In this section, we show the asymptotic properties of the infeasible estimators under the Neyman-Orthogonalized FOC. Then, we introduce the D-RSMD estimator (Debiased Robinson-SMD) as the feasible estimator under the Neyman-Orthogonalized FOC, which shares the same asymptotic properties as the infeasible estimator. The asymptotic properties of the infeasible and feasible estimators under non-orthogonal FOC from Antoine and Sun (2021) are in the Appendix.

3.1 Asymptotic Properties of the Infeasible Estimators

The infeasible estimators under both FOCs have explicit forms and are linear in $\tilde{y}_l$. The SMD estimator introduced in Lavergne and Patilea (2013) has a general form for the asymptotic properties even if there are no explicit forms for these infeasible estimators. Our work here extends Lavergne and Patilea (2013) to allow for nuisance parameters and introduces the debiased method to limit the impact of their estimation.

Proposition 4. *(Consistency and Asymptotic normality of $\tilde{\theta}_{n,o}$)*
Under Assumption 4 and iid assumption for the sample, $\tilde{\theta}_{n,o}$ is consistent for θ_0, that is $\tilde{\theta}_{n,o} \xrightarrow{p} \theta_0$, and asymptotically normally distributed,

$$\sqrt{n}(\tilde{\theta}_{n,o} - \theta_0) \xrightarrow{d} N\left(0, A^{-1}Var[h_1(\tilde{P}_1, \epsilon_1, Z_1, X_1)]\left(A^{-1}\right)'\right)$$

with $A = E\left[\kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)\tilde{P}_l'\right]$

and $\mathtt{Var}\left[h_1(\tilde{P}_j, \epsilon_j, Z_j, X_j)\right] \equiv \mathtt{Var}\left[\int_{\mathbb{R}^{q_z}} e^{-it'Z_j}\epsilon_j\left(E[e^{it'Z_l}\tilde{P}_l] - \frac{g_{0,\tilde{P}_m}(X_j)}{g_{0,\kappa_{m,j}}(X_j)}E[e^{it'Z_l}]\right)d\mu(t)\right]$

$\mathtt{Var}$ stands for the variance-covariance matrix for the vector $h_1(\tilde{P}_1, \epsilon_1, Z_1, X_1)$, which is a conditional mean function defined in Hoeffding (1948b). The expression of $h_1(\tilde{P}_1, \epsilon_1, Z_1, X_1)$ is shown in the proof section of the Appendix. The asymptotic properties of $\tilde{\theta}_{n,o}$ are based on the corresponding properties of U-statistics introduced by Hoeffding (1948b). See Theorem 7.1. The expression for the middle term of the asymptotic variance looks complicated because the $\kappa_{j,l}$ is expressed explicitly. If we show the asymptotic variance with $\kappa_{j,l}$, the form will be simple. The form for the asymptotic variance is shown in the Appendix.

Additionally, we show the similar consistency and asymptotic normality properties of $\tilde{\theta}_{n,u}$ in the Appendix. Because $\tilde{\theta}_{n,u}$ is a direct extension from Antoine and Sun (2021), we only list the expressions and theorems. The detailed proofs for the properties of $\tilde{\theta}_{n,u}$ are shown in the Appendix of Antoine and Sun (2021). The comparison between Neyman orthogonal estimators and non-orthogonal estimators is discussed in the next section.

3.2 Feasible Estimator

The infeasible estimator $\widetilde{\theta}_{n,o}$ depends on unknown nuisance parameters: $g_{0,\tilde{P}_m}(X_l)$, $g_{0,\kappa_{m,l}}(X_l)$, $E(y_l|X_l)$, and $E(P_l|X_l)$. All of the nuisance parameters are conditional expectation functions on covariates. Hence, in practice, we need to find estimators for these conditional expectation functions to obtain the feasible estimator. Replacing every nuisance parameter g_0 with estimators $\hat{g}$ such that $\hat{g}$ converges to g_0 at a rate of $o_p(n^{-1/4})$ will deliver the feasible estimator on θ_0 with $\sqrt{n}$ asymptotic normality.

In this section, we are concerned with models with P variables inside X_l where P can be large. If P is large, we need to assume sparsity for the nuisance parameters, that is, the conditional means can be described with only a few non-zero parameters in front of X_l. The number of non-zero parameters, s_0, is allowed to grow at the rate of $o_p(n^{1/2}/log(P))$. Under the assumed rate for s_0, the Lasso estimation has the following property (see van de Geer (2016)) on the order of mean square error

$$||X(\hat{\beta}-\beta_0)||_2^2/n = \mathcal{O}_p\left(\frac{s_0 log(P)}{n}\right)$$

where $\hat{\beta}$ is the Lasso estimator (Chernozhukov et al. (2018) and van de Geer (2016)) and $||.||_2$ is the $\mathcal{L}_2$ norm. When $3 < P < n$, we can still use the Lasso method with a higher assumed rate for s_0 to select variables. If $P \leq 3$, we can use Nadaraya–Watson estimator to estimate the nuisance parameter with a second degree kernel. $P \leq 3$ is a constraint which is unlikely to hold in practice. Our estimation procedure allows us to handle larger values of P unlike previous literature: for instance, Li and Racine (2007), Robinson (1988), and Antoine and Sun (2021).

After replacing every nuisance parameter with its estimate, we have the feasible estimator $\hat{\theta}_{n,o}$, that is, D-RSMD estimator:

$$\hat{\theta}_{n,o} = \left[\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\widehat{\tilde{P}}_j - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right)\widehat{\tilde{P}_l}'\right]^{-1}\left[\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\widehat{\tilde{P}}_j - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right)\widehat{\tilde{y}}_l\right] \tag{3.10}$$

since $E[\kappa_{j,l}\left(\widehat{\tilde{P}}_j - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right)\widehat{\tilde{P}_l}']$ is invertible.

This leads to the algorithm of our D-RSMD estimation procedure.

Algorithm 3.2. *(Implementation of the D-RSMD estimation procedure)*

(i) conduct Robinson transformation. This step is to estimate the conditional means inside $\widehat{\tilde{P}}_j$ *and* $\widehat{\tilde{y}}_l$*. Any estimators* $\hat{g}$ *with* $\hat{g}$ *converges to* g_0 *at a rate of* $o_p(n^{-1/4})$ *can be applied, for instance, the Lasso method.*

(ii) calculate the $\kappa_{j,l}$*. Inside* $\kappa_{j,l}$*,* $\mu(.)$ *is the CDF of the Gaussian distribution. Because the Fourier transform of the Gaussian is still Gaussian,* $\kappa_{j,l}$ *is easy to compute.*

(iii) *estimate the nuisance parameters* $\widehat{g_{0,\tilde{P}_m}}(X_l)$ *and* $\widehat{g_{0,\kappa_{m,l}}}(X_l)$ *inside the orthogonal FOC. The orthogonal FOC is also orthogonal to the nuisance parameters, so we use the Lasso method to estimate these parameters.*

(iv) *calculate the estimate based on Equation (3.10) for* $\hat{\theta}_{n,o}$*.*

In the approach from the algorithm, all of the nuisance parameters for D-RSMD estimators in the later sections are estimated by the Lasso method with cross validation. The maximum degree of polynomial for the nuisance parameters is 5, which guarantees that the nuisance parameters can be approximated by 5 degree polynomials in all controls and Lasso with cross validation helps us select the controls and their polynomials.

Assumption 5. *$\hat{g}$ converges to g_0 at a rate of $o_p(n^{-1/4})$. For the Lasso method, the number of non-zero parameters s_0 grows at the rate of $o_p(n^{1/2}/log(P))$. For the Nadaraya-Watson estimator, $\sqrt{n}\left(\sum_{s=1}^{q_z} h_s^4 + \left[\frac{1}{nh_1 \ldots h_{q_z}}\right]\right) = o(1)$ where h is the bandwidth.*

Theorem 3.3. *(Consistency and Asymptotic normality of the D-RSMD estimator: $\hat{\theta}_{n,o}$) Under Assumptions 4 - 3 and iid assumption for the sample, $\hat{\theta}_{n,o}$ is consistent, and has an asymptotically normal distribution, that is,*

$$\sqrt{n}(\hat{\theta}_{n,o} - \theta_0) \xrightarrow{d} N\left(0, A^{-1}Var[h_1(\tilde{P}_1, \epsilon_1, Z_1, X_1)]\left(A^{-1}\right)'\right)$$

From Theorem 3.3, the feasible and infeasible estimators share the same asymptotic distribution and are consistent. It is because under Assumption 5, the bias introduced in the first step will not affect the second step substantially. This is the same convergence rate assumption used in Chernozhukov et al. (2018). We expand these properties to the Bierens type estimators.

The asymptotic properties for the feasible version of $\tilde{\theta}_{n,u}$ or the R-SMD estimator from Antoine and Sun (2021) are in the Appendix Section 3.

The asymptotic distributions of the D-RSMD and R-SMD estimators are not the same. When there are many covariates, the D-RSMD estimator using the Lasso for the nuisance parameters works. R-SMD is generated under the condition that the number of covariates (P) is small. When P is small, the D-RSMD estimator is less biased than the R-SMD estimator if they both estimate the nuisance parameters using the same Nadaraya-Watson estimator. There are no analytical results for comparing the asymptotic distributions. In Chernozhukov et al. (2018), the comparison between Neyman orthogonal estimators and non-orthogonal estimators is shown with simulations. We also present the results on bias in Section 4, i.e., the simulation section.

Under heteroskedasticity, the estimator for the variance of the D-RSMD estimator is

$$[n(n-1)C_n]^{-1}\sum_{j=1}^{n}((\sum_{l=1}^{n}\kappa_{j,l}\left(\widehat{\tilde{P}}_l - \frac{\widehat{g_{0,\tilde{P}_l}}(X_j)}{\widehat{g_{0,\kappa_{m,j}}}(X_j)}\right))(\sum_{l=1}^{n}\kappa_{j,l}\left(\widehat{\tilde{P}}_l - \frac{\widehat{g_{0,\tilde{P}_l}}(X_j)}{\widehat{g_{0,\kappa_{m,j}}}(X_j)}\right)')\hat{\epsilon}_j^2)[n(n-1)C_n']^{-1} \quad (3.11)$$

with $C_n = \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{l \neq j}^{n} \kappa_{j,l} \left(\widehat{\tilde{P}}_j - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)} \right) \widehat{\tilde{P}_l}'$

4 Simulation Study

In this section, we conduct 5000 Monte-Carlo replications for two types of Data Generating Process (DGP) to investigate the properties of our D-RSMD estimators for the parameters in the heterogeneous treatment effects of a partially linear model when the number of controls in X_i is greater than or equal to 1. Recall that the partially linear model with heterogeneous treatment effects is in the following.

$$\begin{aligned} y_i &= \theta_{w0} W_i + W_i \cdot X_{i1}' \theta_{wx0} + f_{0,1}(X_i) + \epsilon_i \\ W_i &= I(f_{0,2}(X_i, Z_i) > v_i) \end{aligned} \tag{4.12}$$

$W_i \cdot X_{i1}$ is the interaction term between the treatment and a subvector of covariates. The benchmark model has a nonlinear $f_{0,1}(X_i)$ and a nonlinear function $f_{0,2}(X_i, Z_i)$. $I(.)$ is the indicator function. It will take the value one if the statement inside is true and zero otherwise. In all of the simulations, the key parameters θ_{w0} and θ_{wx0} are 2 and 3 respectively, and we use the same benchmark model.

With the chosen parameter values, the benchmark model is in the following.

$$\begin{aligned} X_i &= X_i^* + 0.4 Z_i \\ W_i &= I(3Z_i + 4Z_i^3 + \sum_{q=1}^{P} \alpha_{1q} X_{qi} + \sum_{q=1}^{P} \alpha_{3q} X_{qi}^3 > -v_i) \\ y_i &= 2W_i + 3W_i X_{1i} + \sum_{q=1}^{P} \beta_{1q} X_{qi} + \sum_{q=1}^{P} \beta_{2q} X_{qi}^2 + \epsilon_i \end{aligned}$$

where $\alpha_{3q} = 2$, $\beta_{2q} = -3$, $\alpha_{1q} = \beta_{1q} = 1$ for $q \leq S$ with S the number of non-zero parameters and P the number of covariates. Because there are nonlinear terms in the model, S is chosen to be a half of s_0, the sparsity level. $\alpha_{3q} = \beta_{2q} = \alpha_{1q} = \beta_{1q} = 0$ when $q > S$.

If β_{2q} is 0 for all q, y_i is linear in X_i. Otherwise, the model is partially linear, because X_i enters nonlinearly. The covariate X_i is the sum of random variables X_i^* generated by a standard normal distribution (or multivariate standard normal distribution) and a fraction of the instrument because our model allows for the correlation between covariate X_i and Z_i. In the benchmark model, we choose the fraction level to be 0.4. The errors (ϵ, v) are bivariate normal distributed with mean 0 and covariance matrix Σ_1 such that $\Sigma_1 = \begin{pmatrix} 1 & 4/9 \\ 4/9 & 1 \end{pmatrix}$. The correlation between ϵ and v makes the treatment W_i an endogenous variable. In the DGP, y_i is not dependent on Z_i directly. The exclusion restrictions assumption is satisfied.

Also, we generate X_i^* and Z_i separately and independently because X_i^* is continuous and Z_i is categorical.

When there is only one control variable X_i, the two nuisance parameters are in the following.

$$\begin{aligned} f_{0,1}(X_i) &= \beta_{11}X_i + \beta_{21}X_i^2 \\ f_{0,2}(X_i, Z_i) &= 3Z_i + 4Z_i^3 + \alpha_{11}X_i + \alpha_{21}X_i^3 \end{aligned}$$

We consider two types of DGPs which mainly differ in how the treatment variable is generated, e.g. using either a categorical instrument with three values, or a binary instrument. Both are motivated by our applications. The first kind of DGP uses a categorical instrument to generate a treatment variable. Specifically, the instrument variable has three values, 0, 1, and 2, and can be interpreted as the sum of two binary variables. This is motivated by our application based on Card (1993) where we construct a similar instrument by adding the indicator of approximation to a four-year college to the indicator for two-year college.

The second type of DGP employs a binary instrument to generate a treatment variable. Correspondingly in Card (1993), this is the indicator of approximation to a four-year college [4]. We generate the binary instrument based on the distribution of the indicator of approximation to a four-year college.

For each DGP, the benchmark case considers 3000 observations and 30 control variables. This is once again in line with both applications. For instance, in Card (1993), there are 3010 valid observations and 27 covariates[5]. We will also consider different sample sizes and a small number of controls.

We report and compare the simulation results for the following estimators.

(i) the D-RSMD estimator proposed in this paper.

(ii) the R-SMD estimator (from Antoine and Sun (2021)) when the number of controls is less than 3.

(iii) the R-GMM estimator which combines Robinson Transformation with GMM. In the Application section, it is called GMM-Lasso, where we use the Lasso method to estimate the nuisance parameters after the Robinson transformation.

(iv) the GMM estimator that treats $f_{0,1}(X_i)$ as a linear function in X_i.

(v) the GMM (Oracle) estimator that uses the true $f_{0,1}(X_i)$.

[4] In the Oregon Health Insurance experiment the lottery variable is also a binary instrument.

[5] Analyzing the Oregon Health insurance experiment, we split the data set into three groups based on age. Each group has about 6000 observations and 21 covariates.

The nuisance parameters in the first step of the Algorithm for the D-RSMD estimator are estimated by the Lasso method because we will use the same estimator when the number of controls is 30. The tuning parameter of the Lasso method (e.g. the penalty term) is selected by cross-validation (see e.g. van de Geer (2016) and Chernozhukov et al. (2018)).

The R-SMD estimator can only be implemented in the first simulation design when the number of control variables is small (e.g. equal to 1). We apply the Nadaraya-Watson estimator, with the rule of thumb bandwidth $h = \sigma_x n^{-0.2}$ to estimate the nuisance parameters in the first step of the R-SMD estimator. D-RSMD and R-SMD estimators both use the CDF of a standard Gaussian distribution as $\mu(.)$ inside $\kappa(Z_j - Z_l) = \int_{R^{q_z}} e^{it'(Z_j - Z_l)} d\mu(t)$. The choice of $\mu(.)$ satisfies Assumption 4(vii).

For each DGP, we report the results for the D-RSMD estimator using only one instrument and two instruments in the following subsection. For the R-SMD estimator, we also report the results for both cases. For GMM type estimators, to estimate two parameters, we need to use at least two instruments. The following is the list of available instrument sets we use for estimators.

(i) Z_1: the binary instrument.

(ii) Z_2: the sum of Z_1 and another binary variable. There are three values in Z_2.

(iii) (Z_1, Z_1X_1): Z_1 and the interaction term between Z_1 and the covariate.

(iv) (Z_2, Z_2X_1): Z_2 and the interaction term between Z_2 and the covariate.

(v) (Z_1, Z_2): Z_1 and Z_2. The correlation between Z_1 and Z_2 is around 0.7.

In all simulation designs, we report the Monte-Carlo Median Bias (Med.Bias), Median Absolute Deviation (MAD), the median of asymptotic standard error under heteroskedasticity (Med.SE), and the Rejection Rate (RR).

4.1 Results for the Model with a Categorical Instrument

In this section, instrument Z is a categorical instrument with three values. That is, for the benchmark model, we generate the treatment by using Z_2. In the simulation, we set the same probability distribution for the instrument Z_2 as that in the data set from Card (1993) to provide insights for the empirical results. Additionally, our notations are consistent throughout the simulation and application parts.

In Table 2.1, we report the simulation results of four D-RSMD, four R-SMD, and six GMM type estimators for θ_{w0} and θ_{wx0}. The sample size is 3,000 and there are 5,000 replications with only one control, that is, $P = 1$. The table contains the results for D-RSMD and R-SMD estimators using one instrument, that is, Z_1 or Z_2, or two instruments, i.e., (Z_1, Z_1X_1) or (Z_2, Z_2X_1). Z_2 is the instrument used in the DGP, so R-SMD using Z_2

or (Z_2, Z_2X_1) will produce better results than it with Z_1 or (Z_1, Z_1X_1). This is also shown in the table. The Nadaraya-Watson estimators for nuisance parameters inside R-SMD and R-GMM estimators when $n = 3000$ and $P = 1$ use a bandwidth of $0.202\hat{\sigma}_x$. GMM type estimators are generated using two instruments (Z_1, Z_1X_1) or (Z_2, Z_2X_1).

		θ_{w0}				θ_{wx0}			
Estimator	*Instrument*	Med.Bias	MAD	Med.SE	RR	Med.Bias	MAD	Med.SE	RR
D-RSMD	Z_1	-0.007	0.069	0.103	0.047	0.019	0.036	0.052	0.067
D-RSMD	Z_2	-0.017	0.129	0.211	0.035	0.034	0.313	0.523	0.035
D-RSMD	(Z_1, Z_1X_1)	0.002	0.077	0.112	0.050	-0.008	0.089	0.130	0.044
D-RSMD	(Z_2, Z_2X_1)	0.001	0.045	0.065	0.048	-0.006	0.055	0.077	0.054
R-SMD	Z_1	-0.207	0.072	0.236	0.007	0.424	0.065	0.449	0.000
R-SMD	Z_2	0.027	0.172	0.280	0.022	-0.060	0.416	0.693	0.018
R-SMD	(Z_1, Z_1X_1)	0.064	0.077	0.127	0.052	-0.112	0.083	0.220	0.003
R-SMD	(Z_2, Z_2X_1)	0.028	0.039	0.072	0.031	-0.053	0.048	0.076	0.079
R-GMM	(Z_1, Z_1X_1)	-0.044	0.076	0.182	0.003	-0.239	0.089	0.245	0.026
R-GMM	(Z_2, Z_2X_1)	-0.015	0.042	0.105	0.001	-0.172	0.053	0.156	0.043
GMM	(Z_1, Z_1X_1)	2.571	0.318	0.472	1.000	-6.003	0.415	0.618	1.000
GMM	(Z_2, Z_2X_1)	2.176	0.231	0.322	1.000	-5.233	0.319	0.454	1.000
GMM (Oracle)	(Z_1, Z_1X_1)	0.000	0.076	0.113	0.048	-0.003	0.093	0.137	0.045
GMM (Oracle)	(Z_2, Z_2X_1)	0.000	0.042	0.061	0.049	0.000	0.053	0.078	0.057

Table 2.1: Categorical instrument when $P = 1$ (n=3000)

Note: Simulation Results for θ_{w0} and θ_{wx0} in the benchmark model using D-RSMD estimator 5000 replications. We report the Monte-Carlo Median Bias (Med.Bias), Median Absolute Deviation (MAD), the median of asymptotic standard error under heteroskedasticity (Med.SE), and Rejection Rate (RR) using a 5% t-test.

Recall that Z_2 is the instrument generating the treatment variable. Intuitively, estimators applying Z_2 or (Z_2, Z_2X_1) will produce better results than those using Z_1 or (Z_1, Z_1X_1) do. Indeed, this is the case for D-RSMD with (Z_2, Z_2X_1), R-SMD, R-GMM, GMM, and GMM (Oracle). From Table 2.1, comparing D-RSMD using (Z_2, Z_2X_1) and GMM (Oracle), we find that all results are close. It suggests that D-RSMD with (Z_2, Z_2X_1) is as good as GMM (Oracle). D-RSMD with (Z_2, Z_2X_1) has a lower median bias than R-SMD with (Z_2, Z_2X_1), which means that the Debiased part in D-RSMD works. Applying the debiased procedure reduces the effect of bias on the estimate introduced by the nuisance parameters. In the table, D-RSMD and R-SMD with Z_2 also work. MAD and Med.SE for D-RSMD are smaller than those for R-SMD. RR for D-RSMD is closer to 5% than R-SMD. The Med.Bias is higher for D-RSMD with Z_2 suggests that the bias introduced by the new nuisance parameters in the debiased procedure has higher effects on the estimate than D-RSMD with (Z_2, Z_2X_1). It is reasonable because the new nuisance parameters in the debiased process inside D-RSMD are the conditional expectations of the $\kappa(Z_j - Z_l)$ where instruments are.

When we use the sets of instruments not in the DGP, e.g., D-RSMD with valid Z_1 or (Z_1, Z_1X_1), the D-RSMD type estimators still work. Some of the results are better than the D-RSMD with Z_2 or (Z_2, Z_2X_1). For instance, The MAD and Med.SE for D-RSMD with Z_1 are smaller than those for D-RSMD with Z_2 for both parameters. This property disappears in the later sections, so it can be case-specific. This also implies that D-RSMD works with various instruments. With valid Z_1 or (Z_1, Z_1X_1), R-SMD or GMM type estimators for θ_{w0} and θ_{wx0} have much higher median biases than the same estimators with valid Z_2 or (Z_2, Z_2X_1). It suggests that choosing the instrument outside of the DGP will cause problems for R-SMD or GMM type estimators.

In the empirical application, there are more than 20 control variables. Hence, the benchmark model in the simulation considers 30 controls to mimic the situation. When there are 30 controls, the controls are generated by a multivariate normal distribution with mean 0 and identity matrix as a covariance matrix. Following the sparsity assumption in Assumption 5, when $P = 30$, we choose sparsity level to be 10, that is, $\alpha_{3q} = 2$, $\beta_{2q} = -3$, $\alpha_{1q} = \beta_{1q} = 1$ when q is less than or equal to 5. We still use the benchmark model in the DGP.

We report the table of robustness check when the model is not sparse, that is, all β_{2q} are not zero, in Supplementary Appendix.

Table 2.2 reports the simulation results of D-RSMD, RSMD, and GMM type estimators for the key parameters θ_{w0} and θ_{wx0} when the sample sizes are 2,000, 3000, or 5000. All of the simulation studies have 5000 replications. The table contains the results for D-RSMD estimators using one instrument Z_1 or Z_2 or two instruments, i.e., (Z_1, Z_1X_1), (Z_2, Z_2X_1), and (Z_1, Z_2). For the GMM type estimators, we report the results using two instruments, that is, (Z_2, Z_2X_1) or (Z_1, Z_2).

When relying on one instrument (e.g., the categorical instrument) in the estimation procedure, only D-RSMD type estimators work. We expect that estimators using Z_2 will have generally better performance than the same estimators using Z_1 in terms of smaller Med.Bias, MAD and Med.SE. RR should be closer to 5% for estimators using Z_2. Indeed, for θ_{w0}, D-RSMD with Z_2 follows this expectation. Although for θ_{w0}, D-RSMD using Z_1 has lower Med.Bias, MAD and Med.SE, the RR is downward bias and not that close to 5%.

When we work with two instruments, i.e., (Z_1, Z_1X_1), (Z_2, Z_2X_1), or (Z_1, Z_2), results for D-RSMD and GMM type estimators are in the 3000 observations panels of Table 2.2. The GMM (Oracle) estimator is a GMM estimator using the correct second degree polynomial for the nonlinear $f_{0,1}(X_i)$. Thus, it has relatively better performance with instrument (Z_2, Z_2X_1), in terms of much smaller bias, lower standard errors and higher t-statistic. It is also shown in the table. However, RR is slightly higher. This higher RR disappears when we have a larger number of observations. See the panel with $n = 5000$.

It is no surprise to see that using another set of instruments (Z_1, Z_2), the GMM (Oracle) estimator does not work so well. It is because the instruments in (Z_1, Z_2) have a higher correlation. If we check the raw estimation results generated by GMM (Oracle) with (Z_1, Z_2), we will see many extreme cases. The same issue happens to GMM as well. However, it is not a big problem for the D-RSMD estimator. D-RSMD estimators using any instrument has relatively stable performance. The only problem with the instrument selection problem

Instrument with 3 values ($P = 30$)

		Estimator	*Ins*	Med.Bias	MAD	Med.SE	RR
$n = 2000$	θ_{w0}	D-RSMD	Z_2	-0.041	0.115	0.146	0.064
		D-RSMD	(Z_2, Z_2X_1)	-0.059	0.119	0.144	0.072
		GMM (Oracle)	(Z_2, Z_2X_1)	-0.006	0.110	0.160	0.050
$n = 2000$	θ_{wx0}	D-RSMD	Z_2	-0.019	0.106	0.158	0.064
		D-RSMD	(Z_2, Z_2X_1)	-0.033	0.079	0.101	0.064
		GMM (Oracle)	(Z_2, Z_2X_1)	0.001	0.059	0.087	0.053
$n = 3000$	θ_{w0}	D-RSMD	Z_1	0.139	0.765	1.140	0.012
		D-RSMD	Z_2	-0.018	0.119	0.158	0.044
		D-RSMD	(Z_1, Z_1X_1)	-0.115	0.557	0.843	0.025
		D-RSMD	(Z_2, Z_2X_1)	-0.028	0.100	0.118	0.055
		D-RSMD	(Z_1, Z_2)	-0.019	0.189	0.267	0.036
		RSMD	Z_1	-0.476	0.773	1.173	0.024
		RSMD	Z_2	-0.134	0.178	0.313	0.083
		RSMD	(Z_1, Z_1X_1)	0.082	0.790	1.158	0.015
		RSMD	(Z_2, Z_2X_1)	-0.202	0.111	0.190	0.157
		RSMD	(Z_1, Z_2)	-0.128	0.267	0.411	0.073
		GMM	(Z_2, Z_2X_1)	3.379	0.897	1.354	0.715
		GMM	(Z_1, Z_2)	12.082	4.798	7.355	0.230
		GMM (Oracle)	(Z_2, Z_2X_1)	0.000	0.089	0.130	0.056
		GMM (Oracle)	(Z_1, Z_2)	0.003	0.506	0.802	0.004
$n = 3000$	θ_{wx0}	D-RSMD	Z_1	-0.013	0.091	0.158	0.020
		D-RSMD	Z_2	-0.019	0.145	0.236	0.048
		D-RSMD	(Z_1, Z_1X_1)	-0.015	0.105	0.154	0.037
		D-RSMD	(Z_2, Z_2X_1)	-0.023	0.068	0.083	0.059
		D-RSMD	(Z_1, Z_2)	-0.018	0.274	0.435	0.034
		RSMD	Z_1	-0.049	0.101	0.161	0.052
		RSMD	Z_2	-0.052	0.231	0.421	0.058
		RSMD	(Z_1, Z_1X_1)	-0.279	0.135	0.260	0.071
		RSMD	(Z_2, Z_2X_1)	-0.074	0.054	0.085	0.129
		RSMD	(Z_1, Z_2)	-0.058	0.374	0.602	0.045
		GMM	(Z_2, Z_2X_1)	-5.957	0.509	0.726	1.000
		GMM	(Z_1, Z_2)	-26.653	9.961	15.090	0.312
		GMM (Oracle)	(Z_2, Z_2X_1)	0.000	0.048	0.071	0.052
		GMM (Oracle)	(Z_1, Z_2)	0.015	1.065	1.712	0.002
$n = 5000$	θ_{w0}	D-RSMD	Z_2	-0.005	0.397	0.694	0.021
		D-RSMD	(Z_2, Z_2X_1)	-0.016	0.082	0.096	0.053
		GMM (Oracle)	(Z_2, Z_2X_1)	-0.001	0.068	0.101	0.050
$n = 5000$	θ_{wx0}	D-RSMD	Z_2	-0.028	0.714	1.267	0.022
		D-RSMD	(Z_2, Z_2X_1)	-0.013	0.054	0.066	0.056
		GMM (Oracle)	(Z_2, Z_2X_1)	-0.001	0.037	0.055	0.049

Table 2.2: Instruments with 3 values

Note: Simulation Results for θ_{w0} and θ_{wx0} in the benchmark model using D-RSMD estimator 5000 replications. We report the Monte-Carlo Median Bias (Med.Bias), Median Absolute Deviation (MAD), median of asymptotic standard error under heteroskedasticity (Med.SE), and Rejection Rate (RR) using a 5% t-test.

happens to the (Z_2, Z_2X_1) where the RR is higher than 5%. This rejection rate decreases when we have a bigger dataset. See the last six rows with $n = 5000$ in Table 2.2.

With a categorical instrument in the DGP, the D-RSMD estimators have the proper size, much smaller bias than GMM estimators. As the sample size grows, the size distortions decrease for (Z_2, Z_2X_1). D-RSMD allows us to use any possible instrument sets to estimate heterogeneous treatment effects. When the correlation between the instruments inside the instrument vector is high, D-RSMD still works, while GMM generates extreme results.

4.2 Results for the Model with a Binary Instrument

The DGP in this section also follows the benchmark model for the categorical instrument variable. The only difference is that when we generate the treatment variable, Z_i is a binary instrument instead of a categorical instrument. Hence, in this section, Z_1 is the true instrument variable. Instrument sets (Z_1, Z_1X_1) and (Z_1, Z_2) contain the true instrument. We report results in Table 2.3.

The first two panels of Table 2.3 contain the preliminary results when $P = 3$. D-RSMD using Z_1 has the lowest MAD and Med.SE among all of the estimators. Its RR are close to 5% for θ_{w0} and slightly oversized for θ_{wx0}. Using the instrument Z_2 (not in the DGP) for D-RSMD will generate higher MAD and Med.SE than utilizing the correct instrument. D-RSMD estimator also works when employing (Z_1, Z_1X_1) and (Z_2, Z_2X_1). When comparing D-RSMD estimators, we see that using the instrument from the DGP or the instrument set including the instrument will generate results with lower MADs and Med.SEs. It is reasonable because working with the correct variable will increase the precision of the estimate. The Med.Bias column shows that for θ_{w0} the D-RSMD estimators using (Z_1, Z_1X_1) and (Z_2, Z_2X_1) have a higher bias than the other D-RSMD estimators. It suggests that using the instrument from the DGP directly will generate a lower bias. This property of D-RSMD also shows in the case where we have 30 controls in the model.

In the simulation, when $P = 30$, we see similar results as the case with a categorical instrument inside DGP. For D-RSMD estimators, using Z_1 or Z_2 already provides close or better results than the GMM (Oracle) in terms of lower MAD and Med.SE. Because the DGP in this section uses the binary instrument, the D-RSMD results with Z_1 should be better than the other D-RSMD estimators in terms of the magnitude of the Med.Bias, MAD, Med.SE. Indeed this is the case. For the RR column, if we compare the Z_1 results with Z_2 ones, we see that the RR is higher in Z_1. It is reasonable because the instrument Z_1 has only two values and Z_2 has three values and Z_2 is the sum of Z_1 and another binary instrument. Z_2 contains more information. The size distortion problem decreases when we generate 5000 observations for each replication.

Comparing the results for instrument sets (Z_1, Z_1X_1) and (Z_1, Z_2), we see that there are similar stories as before. For GMM type estimators, (Z_1, Z_1X_1) is the better choice. For instance, the GMM (Oracle) using (Z_1, Z_1X_1) is the best results for both θ_{w0} and θ_{wx0} with lowest Med.Bias. GMM type estimators using (Z_1, Z_2) still generate scattered results and many extreme cases. The RR for GMM and GMM (Oracle) with (Z_1, Z_2) is close to 0 for θ_{wx0}. Even with the GMM (Oracle) estimator, the correlation between the instruments inside (Z_1, Z_2) still causes GMM estimators a big problem.

If we compare D-RSMD results with GMM (Oracle) estimator, we find that D-RSMD results with Z_1 are close to the Oracle ones for θ_{w0} in terms of MAD, Med.SE, and RR, that suggests that with Z_1 D-RSMD is good enough. For θ_{wx0} estimation, D-RSMD results with Z_1 is better than the GMM (Oracle) using (Z_1, Z_1X_1) in terms of lower MAD and Med.SE. It is reasonable, because we use all of the information from the conditional moment, while GMM (Oracle) only uses two unconditional moments.

Using one binary instrument for D-RSMD generates better results than using two instruments for D-RSMD estimators. It suggests that for D-RSMD, the number of instruments is not a big problem.

5 Empirical Application

5.1 Estimating the Returns of Education on Wages

To illustrate the proposed process, we use the data set from Card (1993) to estimate the heterogeneous returns to education. Many works have used the same data set, for instance, Yanagi (2019), Kitagawa (2015), and Ashenfelter and Rouse (1998). The data is from the National Longitudinal Survey for young men. The detailed information for the variables is in Card (1993). In this paper, we use the same dependent variable, log hourly wages and the same covariates in the baseline model. The education variable in the original work is the years of education. To investigate the treatment effect of college education, we construct the college indicator based on whether the years of education are higher than 14 years. The treatment can be considered as a two-year college degree or higher. It is used in Yanagi (2019) as well. We also conduct the same analysis for years of education. In Kitagawa (2015) the author also treats the education variable as an indicator.

The instrument variable used in Card (1993) is a dummy for growing up near a local four-year college (Z_1). It is used as an instrument because it is not correlated with the individual's ability and increases the probability of attending college. From Kitagawa (2015), the validity of the instrument is not rejected once the covariates are in the estimation. In our models and estimators, all of the covariates are included.

We use the covariates in the original study by Card, for instance, experience, experience squared, age, and so on. In Card (1993), the variables on the family background are an important group of control variables, including the parents' years of education, classes of education, and two indicators for family structure. We incorporate those variables in the analysis. To illustrate the heterogeneous treatment effects of interest, we use the parents' education to generate the interaction term. The parents' education is the average of father's and mother's years of education. A similar control variable is in Ashenfelter and Rouse (1998). The minimum of parents' education is 0. Three individuals' parents have no education. Twenty-four fathers and fifteen mothers have zero years of education. The average value of parents' education is 10.16.

Binary instrument							
		Estimator	*Ins*	Med.Bias	MAD	Med.SE	RR
		D-RSMD	Z_1	-0.006	0.146	0.215	0.048
		D-RSMD	Z_2	-0.004	0.212	0.314	0.043
$n = 3000$	θ_{w0}	D-RSMD	(Z_1, Z_1X_1)	-0.026	0.164	0.247	0.048
$P = 3$		D-RSMD	(Z_2, Z_2X_1)	-0.035	0.230	0.340	0.049
		GMM (Oracle)	(Z_1, Z_1X_1)	-0.004	0.164	0.250	0.048
		D-RSMD	Z_1	-0.005	0.039	0.057	0.052
		D-RSMD	Z_2	-0.004	0.110	0.182	0.050
$n = 3000$	θ_{wx0}	D-RSMD	(Z_1, Z_1X_1)	-0.004	0.124	0.186	0.047
$P = 3$		D-RSMD	(Z_2, Z_2X_1)	-0.001	0.200	0.301	0.040
		GMM (Oracle)	(Z_1, Z_1X_1)	-0.001	0.126	0.191	0.047
		D-RSMD	Z_1	0.028	0.288	0.427	0.045
		D-RSMD	Z_2	0.033	0.524	0.773	0.030
$n = 3000$	θ_{w0}	D-RSMD	(Z_1, Z_1X_1)	-0.259	0.291	0.431	0.082
$P = 30$		D-RSMD	(Z_1, Z_2)	0.041	0.333	0.493	0.041
		RSMD	Z_1	0.128	0.408	0.563	0.072
		RSMD	Z_2	0.154	0.555	0.790	0.039
		RSMD	(Z_1, Z_1X_1)	0.230	0.463	0.572	0.110
		RSMD	(Z_1, Z_2)	0.140	0.438	0.605	0.060
		GMM	(Z_1, Z_1X_1)	7.471	2.968	4.368	0.356
		GMM	(Z_1, Z_2)	6.857	5.691	11.948	0.043
		GMM (Oracle)	(Z_1, Z_1X_1)	0.007	0.266	0.392	0.048
		GMM (Oracle)	(Z_1, Z_2)	-0.011	0.635	1.424	0.003
		D-RSMD	Z_1	-0.012	0.045	0.059	0.064
		D-RSMD	Z_2	-0.012	0.126	0.206	0.037
$n = 3000$	θ_{wx0}	D-RSMD	(Z_1, Z_1X_1)	0.005	0.111	0.163	0.049
$P = 30$		D-RSMD	(Z_1, Z_2)	-0.013	0.181	0.287	0.039
		RSMD	Z_1	-0.020	0.053	0.065	0.105
		RSMD	Z_2	-0.024	0.147	0.302	0.018
		RSMD	(Z_1, Z_1X_1)	-0.337	0.137	0.273	0.144
		RSMD	(Z_1, Z_2)	-0.020	0.195	0.328	0.033
		GMM	(Z_1, Z_1X_1)	-6.735	0.953	1.430	0.999
		GMM	(Z_1, Z_2)	-4.444	21.525	47.007	0.000
		GMM (Oracle)	(Z_1, Z_1X_1)	-0.004	0.107	0.157	0.049
		GMM (Oracle)	(Z_1, Z_2)	0.114	2.502	5.548	0.000
		D-RSMD	Z_1	0.003	0.209	0.314	0.052
$n = 5000$	θ_{w0}	D-RSMD	(Z_1, Z_1X_1)	-0.158	0.210	0.308	0.083
$P = 30$		GMM (Oracle)	(Z_1, Z_1X_1)	0.001	0.203	0.304	0.051
		D-RSMD	Z_1	-0.005	0.029	0.040	0.062
$n = 5000$	θ_{wx0}	D-RSMD	(Z_1, Z_1X_1)	0.006	0.078	0.120	0.045
$P = 30$		GMM (Oracle)	(Z_1, Z_1X_1)	0.002	0.080	0.121	0.045

Table 2.3: Binary Instrument

Note: Simulation Results for θ_{w0} and θ_{wx0} in the benchmark model using D-RSMD estimator 5000 replications. We report the Monte-Carlo Median Bias (Med.Bias), Median Absolute Deviation (MAD), median of asymptotic standard error under heteroskedasticity (Med.SE), and Rejection Rate (RR) using a 5% t-test.

To investigate the properties of different instrument sets, we construct five instrument sets. In three of them, we utilize the information from the dummy of approximation to a two-year college. For instance, Z_2 is the sum of two dummies: growing up near a four-year college and a two-year college. Z_2 generally implies that the more local colleges are, the higher the value is. There are three values, 0, 1, and 2, in Z_2. There are two reasons why Z_2 is generated. The first is that we will use a categorical instrument Z_2 to demonstrate the properties of the new estimator. We also use Z_1 and Z_2 as an instrument set to show that the new method works well with highly correlated instruments. The sets of instruments we considered in this subsection are as follows:

(i) Z_1: the indicator of approximation to a four-year college.

(ii) Z_2: the sum of the indicator of approximation to a two-year college and Z_1. There are 3 values in Z_2.

(iii) (Z_1, Z_1X_1): Z_1 and the interaction term between parents' education and Z_1.

(iv) (Z_1, Z_2): Z_1 and Z_2. The correlation between Z_1 and Z_2 is around 0.729.

We examine four models in the following. Because we have two types of D-RSMD estimators in this section, we denote the one used in simulations as DRSMD-Lasso. The other one is DRSMD-2SOLS.

(i) DRSMD-Lasso estimators:

$$\begin{aligned} y_i &= \theta_{w0} W_i + W_i \cdot X_{i1}' \theta_{wx0} + f_{0,1}(X_i) + \epsilon_i \\ W_i &= I(f_{0,2}(X_i, Z_i) > v_i) \end{aligned}$$

(ii) DRSMD-2SOLS estimators:

$$\begin{aligned} y_i &= \theta_{w0} W_i + W_i \cdot X_{i1}' \theta_{wx0} + \sum_{p=1}^{27} \beta_p X_i + \epsilon_i \\ W_i &= I(f_{0,2}(X_i, Z_i) > v_i) \end{aligned}$$

(iii) GMM estimators:

$$\begin{aligned} y_i &= \theta_{w0} W_i + W_i \cdot X_{i1}' \theta_{wx0} + \sum_{p=1}^{27} \beta_p X_i + \epsilon_i \\ W_i &= I(\alpha_z Z_i + \sum_{p=1}^{27} \alpha_p X_i > v_i) \end{aligned}$$

(iv) GMM-Lasso estimators:

$$\begin{aligned} y_i &= \theta_{w0} W_i + W_i \cdot X_{i1}' \theta_{wx0} + f_{0,1}(X_i) + \epsilon_i \\ W_i &= I(\alpha_z Z_i + \sum_{p=1}^{27} \alpha_p X_i > v_i) \end{aligned}$$

Our new method, the DRSMD-Lasso method allows the covariates X_i to enter the model through an unknown function $f_{0,1}(.)$ and $f_{0,2}(.)$. Because the true function form is unknown, our method will generate more reliable results. The DRSMD-2SOLS estimator allows a nonlinear and unknown $f_{0,2}(.)$, but $f_{0,1}(.)$ needs to be linear. GMM estimator assumes both $f_{0,1}(.)$ and $f_{0,2}(.)$ to be linear in X_i. GMM-Lasso estimator imposes that $f_{0,2}(.)$ is linear. All of the results in the subsection are generated with the college indicator as treatment. [6]

Table 2.4 reports the results for heterogeneous treatment effects of college (W_i). X_{i1} is the parents' education. If we use GMM estimates with (Z_1, Z_1X_1) and (Z_1, Z_2), no results are statistically significant. The GMM results for (Z_1, Z_2) have a large magnitude. There are two moments in (Z_1, Z_2) including the valid instrument. It suggests that using the extra moment is not a good idea. In Table 2.3 of Section 4, the GMM estimators with and without the oracle features using (Z_1, Z_2) also generate unreliable results. In this application, the results for GMM and GMM-Lasso using (Z_1, Z_1X_1) show that the heterogeneous treatment effects are not statistically significant at 5% level. The results from GMM are affected by instruments.

For DRSMD-Lasso estimators, we have obtained different results. DRSMD-Lasso and DRSMD-2SOLS methods with (Z_1, Z_2) as instruments generate more reliable results. Using (Z_1, Z_1X_1) the DRSMD-Lasso and DRSMD-2SOLS produce similar statistically significant results for θ_{w0} and θ_{wx0} at 5% level. The estimate changes across different sets of instruments, but the magnitude does not change too much. It is reasonable because different instruments imply distinct local average treatment effects on the corresponding groups of compliers. When we compare all of the estimates for the overall treatment effect, in Table 2.5, we see that for DRSMD type estimators, the treatment effects are higher than those estimated by GMM type estimators. This difference suggests that the nonlinear part in the $f_{0,2}(.)$ is responsible for the gap between DRSMD and GMM type estimators.

The interpretation of estimates of the parameter θ_{w0} and θ_{wx0} is straightforward. Recall that three individuals' parents have no education. For DRSMD-Lasso with (Z_1, Z_1X_1), when the parents' education is 0, having a college degree increases the average hourly earnings by 299 log points holding other variables constant. When parents' education is 10, having a college degree increases the average hourly earnings by 108 log points. It is not surprising. It means that before 1981, having a college degree or higher almost doubled the average hourly wage ($e^{1.08} - 1$). It is also rare that parents' education is 0 and the child has a college degree. If that is the case, having a college degree helps find a job with higher payments significantly. When we compare the overall effects of heterogeneous treatment effect with

[6] The treatment effects of years of education are reported in Table B.1 in Appendix A. The estimation results for 2SLS from Card (1993) are around 0.132 to 0.140. The DRSMD-Lasso estimator produces similar results. Table B.3 in Appendix A reports heterogeneous treatment effects of years of education.

parents' education being the mean with the homogeneous treatment effect in Appendix B, we find that they are very close, which means that the new method generates reliable results.

Heterogeneous treatment effects				
Estimator for θ_{w0}	Z_1	Z_2	(Z_1, Z_1X_1)	(Z_1, Z_2)
GMM			1.380	-20.348
			(0.858)	(127.277)
GMM-Lasso			-0.821	-13.471
			(1.825)	(50.295)
DRSMD-Lasso	0.731	0.457	2.988**	0.641**
	(0.513)	(0.339)	(1.244)	(0.313)
DRSMD-2SOLS	0.058	-0.505	2.691**	-0.039
	(0.262)	(0.950)	(1.180)	(0.257)
Estimator for θ_{wx0}	Z_1	Z_2	(Z_1, Z_1X_1)	(Z_1, Z_2)
GMM			-0.053	1.882
			(0.058)	(11.288)
GMM-Lasso			0.149	1.320
			(0.148)	(4.700)
DRSMD-Lasso	0.086	0.159	-0.191*	0.065
	(0.078)	(0.173)	(0.099)	(0.052)
DRSMD-2SOLS	0.161	0.302	-0.152**	0.140
	(0.163)	(0.386)	(0.076)	(0.101)

Table 2.4: Heterogeneous Treatment Effects of college education on the log wage

Note: *** Significant at 1%, ** at 5%, * at 10%. θ_{w0} is the parameter in front of the treatment and θ_{wx0} is the parameter for $College \times Parents' education$. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used. Every regression contains 3010 observations.

Average Treatment Effects				
Estimator for $\theta_{w0} + \theta_{wx0}E(X)$	Z_1	Z_2	(Z_1, Z_1X_1)	(Z_1, Z_2)
GMM			0.844**	-1.221
			(0.385)	(12.645)
GMM-Lasso			0.697	-0.054
			(0.711)	(3.192)
DRSMD-Lasso	1.607	2.074	1.042*	1.299*
	(1.261)	(1.924)	(0.628)	(0.727)
DRSMD-2SOLS	1.698	2.560	1.150	1.384*
	(1.45)	(2.998)	(0.749)	(0.828)

Table 2.5: Average Treatment Effects of college education on the log wage

Note: Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used. Every regression contains 3010 observations.

Chapter 3

Estimation of Heterogeneous Treatment Effects: Oregon Health Insurance Experiment

1 Introduction

In early 2008, Oregon used a lottery to select low-income uninsured adults to expand Medicaid enrollment (health coverage). In this process, low-income uninsured adults first register for the waiting list. Then, from the waiting list, the names will be drawn by a lottery. The draw (or the lottery) was random. The randomness provided by the lottery allows researchers to analyze the effects of health insurance on medical, financial, and labour market outcomes. This experiment is a well-known Randomized Control Trial and is studied in many articles, for instance, Baicker et al. (2014) and Finkelstein et al. (2012).

Many studies focus on the treatment effects of health insurance (Medicaid) on health care, employment, debt for health, and many other outcome variables. In those works, the treatment effects are assumed to be homogeneous. In Finkelstein et al. (2012), authors conduct their analysis assuming homogeneity and also use regression analysis to check whether there are heterogeneous treatment effects. They are unable to make precise inferences using traditional estimators like GMM or 2SLS. Studying the heterogeneous treatment effects using only one valid instrument variable needs a more advanced method.

In Chapter 2, the new estimator, the D-RSMD estimator, is generated to estimate the heterogeneous treatment effects using a conditional moment restriction directly. With an instrument variable that satisfies the conditional moment restriction, the estimator uses all of the information from the restriction to allow us to estimate more than one key parameter inside the model. The new estimator is designed for a partially linear model—that is, a model that contains a linear part and a nonparametric part. The key parameters are the parameter in front of the treatment and the parameters in front of the interaction terms inside the linear part of the model. The asymptotic properties of the new estimator are provided in Chapter 2. Also, Chapter 2 includes the simulation results and analysis of the differences between the new estimator and other types of estimators.

In this empirical study, we include the interaction terms inside the model to account for heterogeneity. The new estimator provided in Chapter 2 is used to estimate both parameters within the heterogeneous treatment effects using the only valid instrument variable without generating new variables. The traditional estimation procedure, such as GMM or 2SLS, needs two moments to estimate both parameters within the heterogeneous treatment effects. Without the generated instrument variable, the traditional estimator will face an under-identification problem; that is, one of the key parameters will not be identified. The new estimator with the only valid instrument can identify both. With the new estimation method, we compare the results between the new and traditional estimation methods using or not using the generated instrument variable with reliable inference.

To illustrate the existence of heterogeneous treatment effects in some studies, we revisit the Oregon Health Insurance Experiment and estimate both parameters within the heterogeneous treatment effects to check the impacts of Medicaid from the Oregon Health Insurance Experiment. The public website for this Oregon Medicaid Health Experiment contains all of the related work, data sets, and data descriptions. [1] The original data sets contain several files. In this paper, we use the data set derived from the original data sets. [2]

In the Oregon Health Insurance Experiment, using our estimator with only one valid instrument produces statistically significant results for heterogeneous treatment effects when the GMM estimator does not. Also, if the generated instrument variable is not reliable, using our estimator with only one valid instrument generates more reliable results, and the GMM estimator cannot.

The chapter contributes another empirical study for the new D-RSMD estimator and provides a clear illustration of the advantages of the new estimator in estimates and standard errors. It also contributes to the literature on the Oregon Health Insurance Experiment. With a new estimation procedure, we identify both parameters inside the heterogeneous treatment effects and verify that there are heterogeneous treatment effects, while the traditional estimation method does not.

The paper is organized as follows. Section 2 introduces our data briefly. Section 3 states the framework, motivation, and our estimator and its large sample properties. Section 4 shows the main empirical results for estimating heterogeneous treatment effects. Section 5 contains information on other empirical results. The additional application results are in the Appendix.

[1] See `https://www.nber.org/research/data/oregon-health-insurance-experiment-data`

[2] I would like to thank A. Colin Cameron for the data set and for bringing the Oregon Health Insurance Experiment to my attention.

2 Data

The original data sets are available on the public website, and the detailed description is on the website and in Finkelstein et al. (2012). There are 18572 observations in the dataset we use. Using the dataset, we want to estimate the effects of Medicaid (health coverage) on various outcomes. Because the health conditions, employment, and other outcomes are closely related to the age of the individuals, we believe that age is one of the sources of heterogeneity. We generate the age group variable from the year of the birth variable. First, we calculate the age of each individual. The range of ages is from 21 to 64. Second, we split the original dataset into three or five subsets based on age to create an age group variable. After the split, there are around 5900 to 6900 observations in each subsample. Individuals in the same age group have more similarities. In Section 4, we report the results under three age groups (e.g., 21 to 35, 35 to 50, and 50 to 64). [3]

Additionally, because age is a source of heterogeneity, we can include the age group variable inside the linear part of the partially linear model. The results for this regression framework are provided in Section 5.

There are also many control variables inside the dataset. We use similar controls as used in Baicker et al. (2014) and Finkelstein et al. (2012). These controls include household controls, wave indicators on lottery and survey, and characteristic variables on individuals. We are interested in the heterogeneous treatment effects from the interaction terms generated by an indicator for household income above 50% of the federal poverty line in 2008, household income (hhincome), TANF (cash welfare assistance to low-income families), and cigarette smoking level (smoke). The dependent variables are the current employment indicator, constructed by three indicators on hours of employment (employment), the total out-of-pocket spending on medical care (Out of Pocket Cost), and whether the individual is currently owing money to a health care provider (Debt for Health). All of the dependent variables are from a mail survey starting in July 2009 and ending in March 2010. They are outcomes obtained approximately one year after the treatment. The out-of-pocket cost is in dollars, and hhincome is the household income as a percent of the federal poverty line.

3 Estimators and Framework

Our framework and estimators are based on the estimators in the second chapter. The D-RSMD estimator is an estimation method that combines regularized machine learning methods, Smooth Minimum Distance (SMD) estimation, and Robinson transformation. With the Robinson transformation, the D-RSMD estimator provides the estimation results without assuming the function form for the nonparametric part of the partially linear model.

[3] See Appendix B for tables with five age groups.

For the linear part of the model, we introduce an interaction term between the treatment and a covariate to account for the heterogeneity of the treatment. With the SMD estimation method, the D-RSMD estimator delivers the results of both parameters inside the heterogeneous treatment effects. The parameters we are interested in are the one in front of the treatment Medicaid and the one in front of the interaction term between the treatment and the control variable X_1.

In this model, X_1 represents the vector of control variables that are the sources of the heterogeneity. In Section 4, it is the indicator of the income level, and the estimation results are provided. Section 5 presents the D-RSMD estimation results when X_1 is a vector of control variables. The detailed information will be discussed in the later sections.

To create a better illustration, we compare the D-RSMD estimator and the traditional estimator, for instance, the GMM estimator, in terms of framework. The D-RSMD estimator is reliable in a nonparametric first stage and a partially linear second stage. The framework is in the following:

$$\begin{aligned} Debt_i &= \theta_{w0} Medicaid_i + \theta_{wx0} Medicaid_i \times X_{i1} + f_{0,1}(X_i) + \epsilon_i \\ Medicaid_i &= I(f_{0,2}(X_i, Lottery_i) > v_i) \end{aligned}$$

For one of the traditional estimators, for example, GMM estimators, the framework for the GMM estimator in this section is linear first stage (for the treatment: Medicaid) and linear second stage (for the dependent variable: debt for health). The framework is shown as follows.

$$\begin{aligned} Debt_i &= \theta_{w0} Medicaid_i + \ \theta_{wx0} Medicaid_i \times X_{i1} \ + \ X_i' \beta_x + \epsilon_i \\ Medicaid_i &= I(\alpha_z Lottery_i + X_i' \alpha_x > v_i) \end{aligned}$$

The GMM estimator is not the only estimator that we compare the D-RSMD estimator to. There are other estimators. For instance, the GMM-Lasso estimator combines the Robinson transformation, a regularized machine learning method, and the traditional GMM method. The GMM-Lasso estimator is reliable when the first stage is linear and the second stage is partially linear. The framework for this estimator is as follows.

$$\begin{aligned} Debt_i &= \theta_{w0} Medicaid_i + \ \theta_{wx0} Medicaid_i \times X_{i1} \ + \ f_{0,1}(X_i) + \epsilon_i \\ Medicaid_i &= I(\alpha_z Lottery_i + X_i' \alpha_x > v_i) \end{aligned}$$

The expression of the D-RSMD is shown in Chapter 2. Notice that this is the same estimator as in the previous chapter. In this estimator, the regularized machine learning method we choose to use is still Lasso. Since the D-RSMD estimation method can utilize many kinds of machine learning methods when the convergence rate of these methods satisfy the assumption we provided in the previous chapter, we add "Lasso" to the name of the estimator to indicate we use the Lasso method here. The Asymptotic Properties of the Feasible D-RSMD Estimator have been provided, and the simulation results based on the estimator are listed in the previous chapter. Here, we only provide the expressions for the feasible D-RSMD estimator, the variance estimator of the feasible estimator under heteroskedasticity, and the theorem of the asymptotic properties in Chapter 2 Section 3.

$$\hat{\theta}_{n,o} = \left[\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\widehat{\tilde{P}}_j - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right)\widehat{\tilde{P}_l}'\right]^{-1}\left[\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\widehat{\tilde{P}}_j - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right)\widehat{\tilde{y}}_l\right] \tag{3.1}$$

since $E[\kappa_{j,l}\left(\widehat{\tilde{P}}_j - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right)\widehat{\tilde{P}_l}']$ is invertible.

The Asymptotic Properties of the Feasible D-RSMD Estimator are provided in Theorem 3.3. Here is the Theorem 3.3 in the previous chapter.

Theorem 3.4. *(Consistency and Asymptotic normality of the D-RSMD estimator: $\hat{\theta}_{n,o}$) Under Assumptions 4 - 3 and iid assumption for the sample, $\hat{\theta}_{n,o}$ is consistent, and has an asymptotically normal distribution, that is,*

$$\sqrt{n}(\hat{\theta}_{n,o} - \theta_0) \xrightarrow{d} N\left(0, A^{-1}Var[h_1(\tilde{P}_1, \epsilon_1, Z_1, X_1)]\left(A^{-1}\right)'\right)$$

Under heteroskedasticity, the estimator for variance of the feasible D-RSMD is in the following.

$$[n(n-1)C_n]^{-1}\sum_{j=1}^{n}((\sum_{l=1}^{n}\kappa_{j,l}\left(\widehat{\tilde{P}}_l - \frac{\widehat{g_{0,\tilde{P}_l}}(X_j)}{\widehat{g_{0,\kappa_{l,j}}}(X_j)}\right))(\sum_{l=1}^{n}\kappa_{j,l}\left(\widehat{\tilde{P}}_l - \frac{\widehat{g_{0,\tilde{P}_l}}(X_j)}{\widehat{g_{0,\kappa_{l,j}}}(X_j)}\right)')\hat{\epsilon}_j^2)[n(n-1)C_n']^{-1}$$

with $C_n = \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\ \kappa_{j,l}\left(\widehat{\tilde{P}}_j - \frac{\widehat{g_{0,\tilde{P}_j}}(X_l)}{\widehat{g_{0,\kappa_{j,l}}}(X_l)}\right)\widehat{\tilde{P}_l}'$

The expressions for the D-RSMD estimator and its variance estimator seem complicated in math, but they are easy to compute in matrix form. With the explicit expressions, the computing time will be decreased.

4 Results

In this section, we discuss the possible heterogeneous treatment effects of Medicaid on debt for health when household income is the source of the heterogeneity, after splitting the dataset in 3 subsets by age (see Page 54 for specific details). This procedure allows for straightforward intuition and interpretation. [4] The heterogeneous treatment effects inside the model is presented by the non-zero coefficient for the interaction term between the indicator of household income above 50% federal poverty line and Medicaid. The whole table that contains all of the results for the effects of the treatment (with covariate household income above 50% federal poverty line inside the interaction term) is contained at the end of the section. The additional results for the other dependent variables (or with other covariates inside the interaction term) are in the Appendix. Some necessary robustness checks are contained in Section 5.

We look at two sets of instruments: the lottery (only for D-RSMD) and the lottery and its interaction with X_1 (income greater than 50% of the poverty line). Because D-RSMD is the only estimator that uses the lottery only to estimate both parameters, the D-RSMD results will contain two parts. The first part of the results uses the lottery only, and the second part uses the lottery and its interaction term with X_1. In this way, we can compare the results of the two different instrument sets for the D-RSMD estimation procedure. The GMM and GMM-Lasso estimators can only use the lottery and its interaction term as instruments, which suggests that we can compare the results from D-RSMD and GMM estimators with the lottery and its interaction term as instruments.

Age, in this example, is also considered a source of the heterogeneity. In this section, we split our sample by age into three age groups. In Section 5 we run the regressions for the whole dataset for D-RSMD and GMM estimators, with the age variable agegroup serving as a categorical variable with three values to mimic the three age groups. In this way, we can compare the results in Section 4 and Section 5 to check the properties of estimators with more than one interaction term inside.

The following table (Table 3.1) contains the results for individuals between 36 and 50 years old. In this age group, the total number of observations is 6693. Please keep in mind that X_1 in this table of results stands for "Income above the 50% Federal Poverty Line." Table 3.1 is a subset of Table 3.4. Notice that DRSMD-Lasso is the D-RSMD estimator.

Debt for Health	Estimator for θ_{w0}		Estimator for θ_{wx0}	
	Lottery	(Lottery, Lottery$\times X_{i1}$)	Lottery	(Lottery, Lottery$\times X_{i1}$)
GMM		-0.213*** (0.041)		0.108 (0.085)
DRSMD-Lasso	-0.232*** (0.062)	-0.232*** (0.050)	0.091*** (0.031)	0.088 (0.096)

Table 3.1: results for Individuals Aged 36 to 50

Note: *** Significant at 1%, ** at 5%, * at 10%.

[4] We also consider the full sample (not split by age) and consider possible heterogeneous treatment effects based on income and age: see Page 56 or Section 5.2.

We find heterogeneous treatment effects for individuals between 36 and 50. Estimates obtained by D-RSMD (with the lottery as an instrument or the lottery and its interaction with X_1 as instruments) are very similar. They are also very close to the GMM estimator (with the lottery and its interaction with X_1 as instruments). It suggests that the interaction between the lottery and X_1 is strong and reliable. One important difference between D-RSMD and GMM is that heterogeneous effects are found to be statistically significant only with D-RSMD estimation procedure. In Table 3.1, the age group reports statistically significant results for both θ_{w0} and θ_{wx0} when we use our new estimator with the valid lottery instrument. It suggests the effect of Medicaid on debt for health depends on the income level. Hence, there is heterogeneity.

The interpretation of the estimates is that for people with an income below 50% federal poverty line, Medicaid enrollment decreases their probability of owning money to health providers by 23.2 log points on average, ceteris paribus. For people with an income above 50% federal poverty line, Medicaid enrollment reduces their probability of owning money, but not by that much. It is reasonable because enrollment in Medicaid is not that critical in reducing the debt for individuals with higher incomes compared with individuals with low incomes.

Comparing the results between the two instrument sets (the lottery as an instrument set and the lottery and its interaction with X_1 as the other instrument set), we find that the estimates are close for both parameters, but the standard errors for the θ_{wx0} using the new method are substantially lower, suggesting that using one valid instrument works better. Table 2.3 in the Simulation Section in Chapter 2 also shows that the standard errors of the DRSMD-Lasso estimator for θ_{wx0} are the lowest among all of the estimators.

Table 3.2 has the results for individuals aged 21 to 35. In this age group, the number of observations is 5962. The covariate remains X_1, indicating whether income is greater than 50% of the Federal Poverty Line.

Debt for Health	Estimator for θ_{w0}		Estimator for θ_{wx0}	
	Lottery	(Lottery, Lottery$\times X_{i1}$)	Lottery	(Lottery, Lottery$\times X_{i1}$)
GMM		-0.069 (0.055)		-0.204** (0.096)
DRSMD-Lasso	-0.199*** (0.071)	-0.053 (0.065)	0.058 (0.036)	-0.252** (0.105)

Table 3.2: Results for Individuals Aged 21 to 35

Note: *** Significant at 1%, ** at 5%, * at 10%.

For the age group 21–35, all estimators generate similar results using the second instrument set, that is, the lottery and its interaction term with the indicator for household income. The coefficients for θ_{w0} are not statistically significant, and the ones for θ_{wx0} are statistically significant. It suggests that when households' incomes are higher than 50% federal poverty line, individuals with Medicaid will be less likely to own money to the health providers. It also suggests that Medicaid helps people with higher income levels more than it helps people with lower incomes. Using only the lottery as the instrument, our new procedure generates the opposite results. The interpretation is that for people with lower incomes, Medicaid enrollment decreases their probability of owing money to health providers by 19.9 log points on average, holding other variables constant. For people with higher incomes, the effect of Medicaid decreases.

Based on the D-RSMD results using only the lottery, we do not find support in the data to say that there are heterogeneous treatment effects for individuals between 21 and 35 years old. Furthermore, we discover that the results for individuals aged 21–35 differ significantly between the two instrument sets. It suggests that the generated interaction $Lottery \times X_1$ is invalid.

Next, we re-estimate a homogeneous model. In a homogeneous model, because we only want to estimate the homogeneous treatment effects, the interaction term is not included in the regression model. The traditional estimation method only needs one instrument to estimate one parameter. When we look at the homogeneous treatment effects using both instrument sets in Table 3.3, all estimators generate similar results as the DRSMD-Lasso using only the lottery as the instrument in the heterogeneous treatment effects panel. It also suggests that using a valid instrument to estimate both parameters provides a more reliable outcome. DRSMD-Lasso is reliable under both heterogeneous and homogeneous conditions. Since it is unclear in practice whether the model is homogeneous or not, DRSMD-Lasso appears to be extremely valuable.

Homogeneous Treatment Effects of Medicaid on Debt for Health (Age: 21 - 35)		
Estimator for θ_{w0}	GMM	DRSMD-Lasso
Lottery	-0.161*** (0.049)	-0.172*** (0.063)

Table 3.3: Homogeneous Treatment Effects of Medicaid (Age: 21 - 35)

Note: *** Significant at 1%, ** at 5%, * at 10%.

Table 3.1-3.3 are subsets of Table 3.4. Table 3.4 is provided later. The estimates and standard errors of average treatment effects in Table 3.5 are calculated based on Table 3.4.

5 Robustness Check

In this section, we look into the two different kinds of variations of the previous estimation. In Section 4, the covariate inside the interaction term is the indicator of whether the household income is higher than 50% Federal Poverty Line. To provide more information about these regression results in Section 4, we need to conduct several different robustness checks. In the first subsection, we check the estimation results when the covariate is a dummy. The dummy has a value of 1 when the household income is higher than 100% or 150% Federal

Panel A: Heterogeneous treatment effects						
Debt for Health	Age: 21 - 35		Age: 36 - 50		Age: 51 - 64	
Estimator for θ_{w0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-0.069		-0.213***		-0.193***
		(0.055)		(0.041)		(0.046)
GMM-Lasso		-0.091		-0.259***		-0.245***
		(0.093)		(0.078)		(0.090)
DRSMD-Lasso	-0.199***	-0.053	-0.232***	-0.232***	-0.182**	-0.190***
	(0.071)	(0.065)	(0.062)	(0.050)	(0.072)	(0.056)
Estimator for θ_{wx0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-0.204**		0.108		0.069
		(0.096)		(0.085)		(0.094)
GMM-Lasso		-0.184		0.008		-0.081
		(0.131)		(0.128)		(0.153)
DRSMD-Lasso	0.058	-0.252**	0.091***	0.088	0.076*	0.096
	(0.036)	(0.105)	(0.031)	(0.096)	(0.040)	(0.103)
Panel B: Homogeneous treatment effects						
GMM	-0.161***	-0.130***	-0.170***	-0.189***	-0.164***	-0.177***
	(0.049)	(0.046)	(0.040)	(0.037)	(0.044)	(0.041)
GMM-Lasso	-0.180*	-0.113	-0.256***	-0.260***	-0.279**	-0.244***
	(0.104)	(0.092)	(0.097)	(0.078)	(0.112)	(0.089)
DRSMD-Lasso	-0.172***	-0.187***	-0.198***	-0.198***	-0.150**	-0.145**
	(0.063)	(0.066)	(0.056)	(0.057)	(0.063)	(0.064)
N	5962		6693		5917	

Table 3.4: Heterogeneous Treatment Effects of Medicaid on Debt for Health

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid × Above 50% Federal Poverty Line.

Average Treatment Effects						
Debt for Health	Age: 21 - 35		Age: 36 - 50		Age: 51 - 64	
Estimator for LATE	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-0.190***		-0.150***		-0.149***
		(0.055)		(0.048)		(0.055)
GMM-Lasso		-0.201*		-0.255**		-0.296**
		(0.112)		(0.110)		(0.133)
DRSMD-Lasso	-0.164***	-0.202***	-0.179***	-0.181***	-0.135**	-0.129*
	(0.061)	(0.069)	(0.054)	(0.067)	(0.060)	(0.075)
N	5962		6693		5917	

Table 3.5: Heterogeneous Treatment Effects of Medicaid on Debt for Health

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid × Above 50% Federal Poverty Line. The expression for LATE is $\theta_{w0} + \theta_{wx0}E(X)$.

Poverty Line. In the second subsection, we will not split the dataset into three groups, and we will include the age group variable and its interaction term with the treatment in the model directly. In the last subsection, we will briefly discuss the other robustness checks.

5.1 Estimation Results with Other Covariates

The results are included in Table 3.6 when X_1 is an indicator for "Income Above 100% Federal Poverty Line". Comparing Table 3.6 with Table 3.4, we find that the results are quite similar. For individuals between 36 and 50 years old, there are heterogeneous treatment effects. The estimates show that they benefit more from Medicaid health coverage when they have lower incomes. We do not find support in the data to say that there are heterogeneous treatment effects for individuals between 21 and 35 years old, and the DRSMD-Lasso esimator using only one valid instrument variable generates more reliable results.

Heterogeneous Treatment Effects of Medicaid on Debt for Health				
Debt for Health	Age: 21 - 35		Age: 36 - 50	
Estimator for θ_{w0}	Lottery	(Lottery, Lottery×X_{i1})	Lottery	(Lottery, Lottery×X_{i1})
GMM		-0.123**		-0.179***
		(0.048)		(0.038)
DRSMD-Lasso	-0.178***	-0.123**	-0.202***	-0.200***
	(0.063)	(0.061)	(0.057)	(0.049)
Estimator for θ_{wx0}	Lottery	(Lottery, Lottery×X_{i1})	Lottery	(Lottery, Lottery×X_{i1})
GMM		-0.200		0.072
		(0.148)		(0.159)
DRSMD-Lasso	0.023	-0.212	0.102***	0.032
	(0.033)	(0.154)	(0.038)	(0.192)
N	5962		6693	

Table 3.6: Income above 100% Federal Poverty Line

Note: *** Significant at 1%, ** at 5%, * at 10%.

Table 3.7 reports the outcomes when X_1 is a indicator for "Income Above 150% Federal Poverty Line". Results in Table 3.7 and 3.6 are similar in estimates but different in standard errors. In the table, estimation results from using a dummy for income above 150% Federal Poverty Line show that the treatment effects of Medicaid on debt for people between 36 and 50 are not supported by the data to be heterogeneous, because the estimate for the parameter in front of the interaction term is not statistically significant at the 5% significance level.

The effects of Medicaid on debt for people between 21 and 35 are heterogeneous because the estimate is statistically significant. The difference in the results using two instrument sets for the DRSMD-Lasso estimators suggests that using only the lottery variable as an instrument generates more reliable results. The difference between Table 3.7 and 3.6 shows that the covariate inside the interaction is important for the estimation results. It can be explained by the fact that there are only 808 individuals with incomes above 150% Federal Poverty Line for people between 36 and 50, and 794 individuals for people between 21 and 35 in this data set.

Heterogeneous Treatment Effects of Medicaid on Debt for Health				
Debt for Health	Age: 21 - 35		Age: 36 - 50	
Estimator for θ_{w0}	Lottery	(Lottery, Lottery×X_{i1})	Lottery	(Lottery, Lottery×X_{i1})
GMM		-0.145***		-0.179***
		(0.048)		(0.038)
DRSMD-Lasso	-0.181***	-0.149**	-0.207***	-0.205***
	(0.062)	(0.062)	(0.056)	(0.053)
Estimator for θ_{wx0}	Lottery	(Lottery, Lottery×X_{i1})	Lottery	(Lottery, Lottery×X_{i1})
GMM		-0.204		0.195
		(0.245)		(0.340)
DRSMD-Lasso	0.095*	-0.254	0.067	0.030
	(0.055)	(0.292)	(0.061)	(0.453)
N	5962		6693	

Table 3.7: Income above 150% Federal Poverty Line

Note: *** Significant at 1%, ** at 5%, * at 10%.

5.2 Estimation Results Using the Whole Data Set

This subsection provides the estimation results with an age group variable included in the model, instead of splitting the data set by the age group variable. The difference between the framework in Section 4 and the framework in this subsection lies in the covariate X_1. In this section, X_1 has more than one variable inside. In Section 4, age is also a source of heterogeneity, so the data set is split into 3 groups based on the age of the individuals in the data. In this subsection, through an interaction term, the heterogeneity from age is included in the framework. For the D-RSMD estimation method, including more than two interactions will still work.

The framework in this section also means that we move the X_1 from the nonparametric part of the model to the linear part. Hence, there are three kinds of coefficients in the linear part of the model. The first type of coefficient is the coefficient in front of the treatment. The second type of coefficients is the parameters in front of X_1, such as the parameters measuring the effects of age and income. And the third type of parameters is the parameters in front of the interaction terms. The first and third types of parameters are our key parameters since we want to measure the heterogeneous treatment effects of the treatment.

The lottery is still the instrument in this framework. For the traditional method, the GMM estimators, it uses the instrument set, including lottery, income, age, the interaction term between lottery and income, and the interaction term between lottery and age. That is, to estimate five parameters, the GMM estimators use the instrument set, which includes five instruments, for instance, exogenous variables, lottery, and generated instruments.

The framework for DRSMD-Lasso estimators is summarised in the following equation.

$$
\begin{aligned}
Debt_i = \theta_{w0} Medicaid_i + \theta_{wx0} Medicaid_i \times income_i + \theta_{x10} income_i \\
+ \theta_{x20} agegroup_i + \theta_{x30} Medicaid_i \times agegroup_i + f_{0,1}(X_i) + \epsilon_i
\end{aligned}
$$

$$
Medicaid_i = I(f_{0,2}(X_i, Lottery_i) > v_i)
$$

The framework for GMM estimators includes a linear first stage within the indicator function, as well as a linear second stage for the dependent variable.

$$
\begin{aligned}
Debt_i = \theta_{w0} Medicaid_i + \theta_{wx0} Medicaid_i \times income_i + \theta_{x10} income_i \\
+ \theta_{x20} agegroup_i + \theta_{x30} Medicaid_i \times agegroup_i + X_i' \beta_x + \epsilon_i
\end{aligned}
$$

$$
Medicaid_i = I(\alpha_z Lottery_i + X_i' \alpha_x > v_i)
$$

The model for GMM-Lasso estimators includes a linear first stage for the indicator function and a partially-linear second stage for the dependent variable.

$$
\begin{aligned}
Debt_i = \theta_{w0} Medicaid_i + \theta_{wx0} Medicaid_i \times income_i + \theta_{x10} income_i \\
+ \theta_{x20} agegroup_i + \theta_{x30} Medicaid_i \times agegroup_i + f_{0,1}(X_i) + \epsilon_i
\end{aligned}
$$

$$
Medicaid_i = I(\alpha_z Lottery_i + X_i' \alpha_x > v_i)
$$

In Table 3.8, we show the results for three estimators from the regressions when we extract income outside the nonparametric part of the model. For DRSMD-Lasso estimator, the estimates for Medicaid, the interaction between Medicaid and income are statistically significant at 5% significance level when using only one instrument, and the estimates for income and the interaction term between Medicaid and age are statistically significant at 15% significance level when using only one instrument. These results allow us to draw similar conclusions as in Section 4. That is, using only the valid instrument, the model shows that the treatment of Medicaid is heterogeneous, and the heterogeneity comes from age and income. After analyzing the outcomes from the GMM estimator or the GMM-Lasso using the instrument set (Z_1, Z_1X_1), we would incorrectly conclude that there is no heterogeneity in the treatment effects. This shows that the generated instruments are problematic.

5.3 Other Robustness Checks

For other robustness check, we report the results when the covariate X_1 is TANF in the appendix. TANF variable is known as a variable with little variation. For instance, there

Heterogeneous treatment effects						
Debt for Health	GMM		GMM-Lasso		DRSMD-Lasso	
Variables	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
Medicaid		-0.148***		-0.206*	-0.187***	-0.178***
		(0.044)		(0.115)	(0.037)	(0.046)
Medicaid*income		0.003		-0.045	0.074***	-0.018
		(0.053)		(0.068)	(0.020)	(0.056)
income		0.003		0.014	-0.025.	0.005
		(0.020)		(0.062)	(0.016)	(0.025)
agegroup		-0.007		-0.010	-0.006	-0.007
		(0.010)		(0.021)	(0.005)	(0.009)
Medicaid*agegroup		-0.019		-0.016	-0.018.	-0.016
		(0.032)		(0.054)	(0.011)	(0.028)

Table 3.8: Heterogeneous Treatment Effects for Robustness Check (5 parameters)

Note: *** Significant at 1%, ** at 5%, * at 10%, . at 15%.

are only 2% of individuals on TANF. We want to see whether this will affect our results. Table C.5 reports the case where y_i is still Debt for Health and there are five age groups. The DRSMD-Lasso with Z_1 estimate for θ_{w0} is -0.215 and for θ_{wx0} is 0.307. Both of them are statistically significant at 5% level. It suggests that for an individual between 21 and 29 years old, TANF will decrease the negative effect of Medicaid on the probability of owning money. We also see that the new method using the valid instrument Z_1 still generates reliable results with the lowest standard error.

Table C.1 contains the results after splitting the sample into 5 age groups. Table C.11 and Table C.21 report the results for the effects of treatment on Employment and Out of Pocket Cost correspondingly. Both of the two tables show that there are heterogeneous treatment effects for young individuals (between 21 and 29 years old) when X_1 is a dummy for "Income Above 50% Federal Poverty Line".

Bibliography

Ai, C. and X. Chen (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica 71*, 1795–1843.

Angrist, J., K. Graddy, and G. Imbens (2000). The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *The Review of Economic Studies 67*, 499–527.

Antoine, B. and P. Lavergne (2014). Conditional Moment Models under Semi-strong Identification. *Journal of Econometrics 182*(1), 59 – 69. Causality, Prediction, and Specification Analysis: Recent Advances and Future Directions.

Antoine, B. and P. Lavergne (2020). Identification-robust Non-parametric Inference in a Linear IV model. *Working paper*. https://ideas.repec.org/p/sfu/sfudps/dp20-03.html.

Antoine, B. and X. Sun (2021). Partially linear models with endogeneity: a conditional moment-based approach. *The econometrics journal 25*(1), 256–275.

Ashenfelter, O. and C. Rouse (1998). Income, schooling, and ability: Evidence from a new sample of identical twins. *The Quarterly journal of economics 113*(1), 253–284.

Baicker, K., A. Finkelstein, J. Song, and S. Taubman (2014). The impact of medicaid on labor market activity and program participation: Evidence from the oregon health insurance experiment. *The American economic review 104*(5), 322–328.

Bekker, P. (1994). Alternative approximations to the distributions of instrumental variables estimators. *Econometrica 62*, 657–681.

Bierens, H. (1982). Consistent model specification tests. *Journal of Econometrics 20*(1), 105–134.

Bierens, H. (1990). A Consistent Conditional Moment Test of Functional Form. *Econometrica 58*, 1443–1458.

Bierens, H. and W. Ploberger (1997). Asymptotic Theory of Integrated Conditional Moment Tests. *Econometrica 65*, 1129–1151.

Bose, A. and S. Chatterjee (2018). *U-Statistics, Mm-Estimators and Resampling.* Springer Singapore.

Breunig, C. and X. Chen (2020). Adaptive, rate-optimal testing in instrumental variables models. *Working paper*. https://arxiv.org/abs/2006.09587.

Card, D. (1993). Using geographic variation in college proximity to estimate the return to schooling. Working Paper 4483, National Bureau of Economic Research.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018a). Alternative Asymptotics and the Partially Linear Model With Many Regressors. *Econometric Theory 34*, 277–301.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018b). Inference in Linear Regression Models with Many Covariates and Heteroscedasticity. *Journal of the American Statistical Association 113*, 1350–1361.

Champeney, D. C. (1987). *A Handbook of Fourier Theorems*. Cambridge University Press.

Chao, J., N. Swanson, J. Hausman, W. Newey, and T. Woutersen (2012). Asymptotic distribution of JIVE in a heteroskedastic IV regression with many instruments. *Econometric Theory 28*, 42–86.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The econometrics journal 21*(1), C1–C68.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. K. Newey (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*, C1–C68.

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins (2018). Locally robust semiparametric estimation. *Working paper*. https://arxiv.org/abs/1608.00033.

Dieterle, S. G. and A. Snell (2016). A simple diagnostic to investigate instrument validity and heterogeneous effects when using a single instrument. *Labour Economics 42*(C), 76–86.

Dinkelman, T. (2011). The effects of rural electrification on employment: new evidence from South Africa. *American Economic Review 101*, 3078–3108.

Dominguez, M. and I. Lobato (2004). Consistent estimation of models defined by conditional moment restrictions. *Econometrica 72*(5), 1601–1615.

Engle, R., C. Granger, J. Rice, and A. Weiss (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association 81*, 310–320.

Escanciano, J. (2018). A simple and robust estimator for linear regression models with strictly exogenous instruments. *The Econometric Journal 21*, 36–54.

Finkelstein, A., S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen, and K. Baicker (2012). The oregon health insurance experiment: Evidence from the first year. *The Quarterly journal of economics 127*(3), 1057–1106.

Härdle, W., H. Liang, and J. Gao (2000). *Partially Linear Models*. Physica-Verlag.

Hausman, J., W. K. Newey, T. Woutersen, J. C. Chao, and N. R. Swanson (2012). Instrumental Variable Estimation with Heteroskedasticity and Many Instruments. *Quantitative Economics 3*, 211–255.

Heckman, J., S. Urzua, and E. Vytlacil (2006). Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics 88*, 389–432.

Hoeffding, W. (1948a). A Class of Statistics with Asymptotically Normal Distribution. *Annals of Mathematical Statistics 19*, 293–325.

Hoeffding, W. (1948b). A class of statistics with asymptotically normal distribution. *The Annals of mathematical statistics 19*(3), 293–325.

Hurst, S. (1995). *The characteristic function of the student t distribution.* Canberra : Centre for Mathematics and its Applications, School of Mathematical Sciences, ANU.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press.

Johnson, N., S. Kotz, and N. Balakrishnan (1995). *Continuous Univariate Distributions* (2nd ed. ed.). Wiley, New York.

Jun, S. and J. Pinkse (2012). Testing Under Weak Identification with Conditional Moment Restrictions. *Econometric Theory 28*, 1229–1282.

Kitagawa, T. (2015). A test for instrument validity. *Econometrica 83*(5), 2043–2063.

Lavergne, P. and V. Patilea (2013). Smooth minimum distance estimation and testing with conditional estimating equations: Uniform in bandwidth theory. *Journal of Econometrics 177*, 47–59.

Li, Q. and J. S. Racine (2007). *Nonparametric Econometrics: Theory and Practice.* Princeton University Press.

Nekipelov, D., V. Semenova, and V. Syrgkanis (2018). Regularized orthogonal machine learning for nonlinear semiparametric models.

Otsu, T. (2011). Empirical Likelihood Estimation of Conditional Moment Restriction Models with unknown functions. *Econometric Theory 27*, 8–46.

Robinson, P. (1988). Root-n-consistent semiparametric regression. *Econometrica 56*(4), 931–954.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics.* Wiley series in probability and mathematical statistics. John Wiley and Sons, Inc.

Stinchcombe, M. and H. White (1998). Consistent Specification Testing With Nuisance Parameters Present Only Under the Alternative. *Econometric Theory 14*, 295–325.

van de Geer, S. (2016). *Estimation and Testing Under Sparsity École d'Été de Probabilités de Saint-Flour XLV – 2015 / by Sara van de Geer.* (1st ed. 2016. ed.). École d'Été de Probabilités de Saint-Flour, 2159. Springer International Publishing : Imprint: Springer.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data / Jeffrey M. Wooldridge.* (2nd ed. ed.). MIT Press.

Xu, R. (2019). Weak Instruments with a Binary Endogenous Explanatory Variable. *Working paper*. https://xuruonan.weebly.com/research.html.

Yanagi, T. (2019). Inference on local average treatment effects for misclassified treatment. *Econometric reviews 38*(8), 938–960.

Appendix A

Partially Linear Models with Endogeneity: a conditional moment based approach

1 Proofs of the main theoretical results

1.1 Equivalence between the objective functions (2.5) and (2.6)

Proof. The objective function (2.5) can be written as

$$M_\infty(\beta) = \int_{\mathbb{R}^{q_w}} E(e_j(\beta)e^{it'W_j})E(e_l(\beta)e^{-it'W_l})d\mu(t)$$

From Assumption 1(v), for all $j \neq l$, $Cov(e_j(\beta)e^{it'W_j}, e_l(\beta)e^{-it'W_l}) = 0$. Thus, for all $j \neq l$, we have:

$$\begin{aligned} M_\infty(\beta) &= \int_{\mathbb{R}^{q_w}} E(e_j(\beta)e^{it'W_j}e_l(\beta)e^{-it'W_l})d\mu(t) \\ &= \int_{\mathbb{R}^{q_w}} E(e_j(\beta)e_l(\beta)e^{it'(W_j-W_l)})d\mu(t) \\ &= E(\int_{\mathbb{R}^{q_w}} e_j(\beta)e_l(\beta)e^{it'(W_j-W_l)}d\mu(t)) \end{aligned}$$

Thus, the objective function becomes

$$M_\infty(\beta) = E(e_j(\beta)e_l(\beta)\kappa_{j,l})$$

where $\kappa_{j,l} = k(W_j - W_l) = \int_{\mathbb{R}^{q_w}} e^{it'(W_j-W_l)}d\mu(t)$. And $k(u)$ is the inverse Fourier transform of $d\mu(t)$ with $u = W_j - W_l$. □

1.2 Proof of Proposition 1

Proof. From (2.5), we have: $M_\infty(\beta) \geq 0$ and $M_\infty(\beta_0) = 0$ since $E(e_j(\beta_0)|W_j) = 0$ for any j.
Since (2.5) can be written as (2.6), as long as $j \neq l$, β_0 minimizes $E(e_j(\beta)e_l(\beta)\kappa_{j,l})$, and $E(e_j(\beta_0)e_l(\beta_0)\kappa_{j,l}) = 0$. The associated FOC write:

$$\begin{aligned}
& E(\tilde{X}_j(\tilde{y}_l - \tilde{X}_l'\beta)\kappa_{j,l} + (\tilde{y}_j - \tilde{X}_j'\beta)\tilde{X}_l\kappa_{j,l}) = 0 \\
\Rightarrow \quad & E(\tilde{X}_j(\tilde{y}_l - \tilde{X}_l'\beta)\kappa_{j,l}) + E(\tilde{y}_j - \tilde{X}_j'\beta)\tilde{X}_l\kappa_{j,l}) = 0 \\
\Rightarrow \quad & E(\tilde{X}_j(\tilde{y}_l - \tilde{X}_l'\beta)\kappa_{j,l}) = 0
\end{aligned}$$

since $E(\tilde{X}_j(\tilde{y}_l - \tilde{X}_l'\beta)\kappa_{j,l}) = 0$ under Assumption 1(vi). Hence we have:

$$E(\kappa_{j,l}\tilde{X}_j\tilde{y}_l - \kappa_{j,l}\tilde{X}_j\tilde{X}_l'\beta) = 0$$

and provided $E(\kappa_{j,l}\tilde{X}_j\tilde{X}_l')$ is nonsingular, we have a unique minimizer,

$$\beta_0 = [E(\kappa_{j,l}\tilde{X}_j\tilde{X}_l')]^{-1}E(\kappa_{j,l}\tilde{X}_j\tilde{y}_l)$$

To show that $E(\kappa_{j,l}\tilde{X}_j\tilde{X}_l')$ is nonsingular, we consider the associated quadratic form, and show that it is positive definite. For any a real vector of size p, we have:

$$\begin{aligned}
E(a'\tilde{X}_j\tilde{X}_l'a\kappa_{j,l}) &= E(\kappa_{j,l}a'E(\tilde{X}_j|W_j)E(\tilde{X}_l'a|W_l)) \\
&= E(\int_{R^{q_w}} e^{it'(W_j-W_l)}d\mu(t)a'E(\tilde{X}_j|W_j)E(\tilde{X}_l'|W_l)a) \\
&= \int_{R^{q_w}} E[e^{it'(W_j-W_l)}a'E(\tilde{X}_j|W_j)E(\tilde{X}_l'|W_l)a]d\mu(t) \\
&= \int_{R^{q_w}} E[e^{it'W_j}a'E(\tilde{X}_j|W_j)E(\tilde{X}_l'|W_l)ae^{-it'W_l}]d\mu(t) \\
&= \int_{R^{q_w}} E[a'e^{it'W_j}E(\tilde{X}_j|W_j)]E[E(\tilde{X}_l'|W_l)ae^{-it'W_l}]d\mu(t) \\
&= \int_{R^{q_w}} E[a'e^{it'W_j}E(\tilde{X}_j|W_j)]E[E(\tilde{X}_j'|W_j)ae^{-it'W_j}]d\mu(t) \\
&= \int_{R^{q_w}} |\left(\int_{R^{q_w}} a'e^{it'W_j}E(\tilde{X}_j|W_j)f_W(W_j)dW_j\right)|^2 d\mu(t) \\
&= (2\pi)^{2q_w}\int_{R^{q_w}} |\left(\mathcal{F}[a'E(\tilde{X}_j|W_j)f_W(W_j)](t)\right)|^2 d\mu(t) \\
&\geq 0
\end{aligned}$$

with μ strictly positive on $\mathbb{R}^p$ and $\mathcal{F}[g]$ the Fourier transform of a well-defined function $g(.)$ on $\mathbb{R}^{q_w}$ formally defined as,

$$\mathcal{F}[g](t) = \frac{1}{(2\pi)^{q_w}}\int \exp^{it'u} g(u)du\,. \tag{1.1}$$

We then have:

$$\begin{aligned} a'E(\kappa_{j,l}\tilde{X}_j\tilde{X}_l')a = 0 \quad &\Leftrightarrow \quad \exists\ a \neq 0\ s.t.\ a'E(\tilde{X}_j|W_j)f(W_j) = 0\ a.s. \\ &\Leftrightarrow \quad \exists\ a \neq 0\ s.t.\ a'E(\tilde{X}_j|W_j) = 0\ a.s. \end{aligned}$$

This cannot hold, since, by Assumptions $1(ii)$ and $1(iii)$, $E(\tilde{X}_j|W_j) \neq 0$ a.s. and $E(\tilde{X}_j\tilde{X}_j')$ is nonsingular. □

1.3 Closed-form expressions for $\tilde{\beta}_n$ and $\hat{\beta}_n$

- Closed-form expression for $\tilde{\beta}_n$:

Proof. From (2.7), the associated FOC write:

$$\begin{aligned} \tilde{X}'\tilde{\kappa}[\tilde{y} - \tilde{X}\tilde{\beta}] = 0 \quad &\Rightarrow \quad \tilde{X}'\tilde{\kappa}\tilde{y} - \tilde{X}'\tilde{\kappa}\tilde{X}\tilde{\beta} = 0 \\ &\Rightarrow \quad \tilde{X}'\tilde{\kappa}\tilde{y} = \tilde{X}'\tilde{\kappa}\tilde{X}\tilde{\beta} \end{aligned}$$

Since $E(k(W_j - W_l)\tilde{X}_j\tilde{X}_l')$ is nonsingular, $\tilde{X}'\tilde{\kappa}\tilde{X}$ is also invertible for n large enough, and we have:

$$\tilde{\beta} = [\tilde{X}'\tilde{\kappa}\tilde{X}]^{-1}\tilde{X}'\tilde{\kappa}\tilde{y}$$

□

- Closed-form expression for $\hat{\beta}_n$:

Proof. From (3.9), the FOC are:

$$\begin{aligned} \widehat{\tilde{X}}'\tilde{\kappa}[\widehat{\tilde{y}} - \widehat{\tilde{X}}\hat{\beta}] = 0 \quad &\Rightarrow \quad \widehat{\tilde{X}}'\tilde{\kappa}\widehat{\tilde{y}} - \widehat{\tilde{X}}'\tilde{\kappa}\widehat{\tilde{X}}\hat{\beta} = 0 \\ &\Rightarrow \quad \widehat{\tilde{X}}'\tilde{\kappa}\widehat{\tilde{y}} = \widehat{\tilde{X}}'\tilde{\kappa}\widehat{\tilde{X}}\hat{\beta} \\ &\Rightarrow \quad \hat{\beta} = [\widehat{\tilde{X}}'\tilde{\kappa}\widehat{\tilde{X}}]^{-1}\widehat{\tilde{X}}'\tilde{\kappa}\widehat{\tilde{y}} \end{aligned}$$

since $E[\widehat{\tilde{X}}'\tilde{\kappa}\widehat{\tilde{X}}]$ is invertible. □

1.4 Proof of Proposition 2:

Proof.

$$\begin{aligned}
\tilde{\beta} &= [\tilde{X}'\tilde{\kappa}\tilde{X}]^{-1}\tilde{X}'\tilde{\kappa}\tilde{y} \\
&= [\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{X}_j\tilde{X}_l']^{-1}[\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{X}_j\tilde{y}_l] \\
&= [\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{X}_j\tilde{X}_l']^{-1}[\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{X}_j(\tilde{X}_l'\beta_0+e_l)] \\
&= [\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{X}_j\tilde{X}_l']^{-1}[\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{X}_j\tilde{X}_l'\beta_0+\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{X}_je_l] \\
&= [\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{X}_j\tilde{X}_l']^{-1}[\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{X}_j\tilde{X}_l'\beta_0]+[\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{X}_j\tilde{X}_l']^{-1}[\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{X}_je_l] \\
&= \beta_0+\left[\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{X}_j\tilde{X}_l'\right]^{-1}\left[\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{X}_je_l\right]
\end{aligned}$$

Denote $A_n=\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{X}_j\tilde{X}_l'$ and $B_n=\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{X}_je_l$. The remainder of the proof is organized in 3 steps: (i) we first show that A_n is a U-statistic and find its probability limit by applying a WLLN for U-statistics; (ii) we then show that B_n is also a U-statistic and find its probability limit by applying a WLLN for U-statistics and its asymptotic distribution by applying a CLT for U-statistics; (iii) we conclude the proof by showing that the necessary WLLN and CLT for U-statistics apply under Assumption 2.

(i) To show that A_n is a U-statistic, notice that

$$\begin{aligned}
A_n &= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{X}_j\tilde{X}_l' \\
&= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{j<l}^{n}\kappa_{j,l}\tilde{X}_j\tilde{X}_l'+\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{j>l}^{n}\kappa_{j,l}\tilde{X}_j\tilde{X}_l' \\
&= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{j<l}^{n}\kappa_{j,l}\tilde{X}_j\tilde{X}_l'+\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{j<l}^{n}\kappa_{l,j}\tilde{X}_l\tilde{X}_j' \\
&= \frac{1}{2}\frac{2}{n(n-1)}\sum_{j=1}^{n}\sum_{j<l}^{n}\kappa_{j,l}\tilde{X}_j\tilde{X}_l'+\frac{1}{2}\frac{2}{n(n-1)}\sum_{j=1}^{n}\sum_{j<l}^{n}\kappa_{l,j}\tilde{X}_l\tilde{X}_j' \\
&= \frac{1}{2}\frac{2}{n(n-1)}\sum_{j=1}^{n}\sum_{j<l}^{n}(\kappa_{j,l}\tilde{X}_j\tilde{X}_l'+\kappa_{l,j}\tilde{X}_l\tilde{X}_j')
\end{aligned}$$

Hence, A_n is a half of a U-statistics, and, under Assumption 2, a WLLN for U-statistics applies to get:

$$A_n\xrightarrow{p}A \qquad \text{with} \qquad A\equiv\frac{1}{2}E\left[\frac{2}{n(n-1)}\sum_{j=1}^{n}\sum_{j<l}^{n}(\kappa_{j,l}\tilde{X}_j\tilde{X}_l'+\kappa_{l,j}\tilde{X}_l\tilde{X}_j')\right]=E(\kappa_{j,l}\tilde{X}_j\tilde{X}_l')$$

since

$$A = \frac{1}{2}E(\kappa_{j,l}\tilde{X}_j\tilde{X}_l' + \kappa_{l,j}\tilde{X}_l\tilde{X}_j') = \frac{1}{2}E(\kappa_{j,l}\tilde{X}_j\tilde{X}_l') + \frac{1}{2}E(\kappa_{l,j}\tilde{X}_l\tilde{X}_j') = E(\kappa_{j,l}\tilde{X}_j\tilde{X}_l')$$

with A nonsingular under Assumption 1 (as show previously).

(ii) To show that B_n is a U-statistic, notice that:

$$B_n = \frac{1}{n(n-1)}\sum_{j<l}^{n}(\kappa_{j,l}\tilde{X}_je_l + \kappa_{l,j}\tilde{X}_le_j)$$

Define $h(\tilde{x}_1, \epsilon_1, w_1; \tilde{x}_2, \epsilon_2, w_2) = \kappa_{1,2}\tilde{x}_1\epsilon_2 + \kappa_{2,1}\tilde{x}_2\epsilon_1$. Since h is a symmetric function of observations 1 and 2, a U-statistic with kernel h is defined as

$$B_n' = \frac{2}{n(n-1)}\sum_{j<l}^{n}h(\tilde{X}_j, e_j, W_j; \tilde{X}_l, e_l, W_l) \qquad \text{and} \qquad B_n = \frac{1}{2}B_n'$$

And we have:

$$\begin{aligned} E(B_n') &= E(\kappa_{j,l}\tilde{X}_je_l + \kappa_{l,j}\tilde{X}_le_j) \\ &= 2E(\kappa_{j,l}\tilde{X}_je_l) \\ &= 2\int_{\mathbb{R}^{q_w}} E[\tilde{X}_je_le^{it'(W_j-W_l)}]d\mu(t) \\ &= 2\int_{\mathbb{R}^{q_w}} E[\tilde{X}_je^{it'W_j}e_le^{-it'W_l}]d\mu(t) \\ &= 2\int_{\mathbb{R}^{q_w}} E[\tilde{X}_je^{it'W_j}]E[e_le^{-it'W_l}]d\mu(t) \\ &= 0 \qquad \text{since } E[e_le^{-it'W_l}] = 0. \end{aligned}$$

Hence, $E(B_n) = 0$. According to WLLN for U-statistics, we have $B_n \xrightarrow{p} 0$, and we conclude that $\tilde{\beta}_n$ is a consistent estimator of β_0.

To derive the asymptotic normality, we first need to compute the asymptotic variance for the U-statistic B_n', which means that we need to find the variance for

$$E(h(\tilde{X}_1, e_1, W_1; \tilde{X}_2, e_2, W_2)|\tilde{X}_1 = \tilde{x}_1, e_1 = \epsilon_1, W_1 = w_1)\,.$$

Let $h_1(\tilde{x}_1, \epsilon_1, w_1) \equiv E(h(\tilde{X}_1, e_1, W_1; \tilde{X}_2, e_2, W_2)|\tilde{X}_1 = \tilde{x}_1, e_1 = \epsilon_1, W_1 = w_1)$. We have:

$$h(\tilde{x}_1, \epsilon_1, w_1; \tilde{x}_2, \epsilon_2, w_2) \;=\; \kappa_{1,2}\tilde{x}_1\epsilon_2 + \kappa_{2,1}\tilde{x}_2\epsilon_1$$

and

$$\begin{aligned} h_1(\tilde{x}_1, \epsilon_1, w_1) &= E[\int_{\mathbb{R}^{q_w}} e^{it'(w_1-W_2)}d\mu(t)\tilde{x}_1e_2 + \int_{R^{q_w}} e^{it'(W_2-w_1)}d\mu(t)\tilde{X}_2\epsilon_1] \\ &= E[\int_{\mathbb{R}^{q_w}} e^{it'(w_1-W_2)}d\mu(t)\tilde{x}_1e_2] + E[\int_{R^{q_w}} e^{it'(W_2-w_1)}d\mu(t)\tilde{X}_2\epsilon_1] \end{aligned}$$

The first element of the right hand side is

$$\begin{aligned} E[\int_{\mathbb{R}^{qw}} e^{it'(w_1-W_2)}d\mu(t)\tilde{x}_1e_2] &= \int_{\mathbb{R}^{qw}} E[e^{it'w_1}e^{-it'W_2}\tilde{x}_1e_2]d\mu(t) \\ &= \int_{\mathbb{R}^{qw}} e^{it'w_1}\tilde{x}_1E[e^{-it'W_2}e_2]d\mu(t) = 0 \end{aligned}$$

The second term of the right hand side is

$$\begin{aligned} E[\int_{\mathbb{R}^{qw}} e^{it'(W_2-w_1)}d\mu(t)\tilde{X}_2\epsilon_1] &= \int_{\mathbb{R}^{qw}} E[e^{it'W_2}e^{-it'w_1}\tilde{X}_2\epsilon_1]d\mu(t) \\ &= \int_{\mathbb{R}^{qw}} e^{-it'w_1}\epsilon_1E[e^{it'W_2}\tilde{X}_2]d\mu(t) \end{aligned}$$

Hence, $h_1(\tilde{x}_1,\epsilon_1,w_1) = \int_{R^{qw}} e^{-it'w_1}\epsilon_1E[e^{it'W_2}\tilde{X}_2]d\mu(t)$. Since $E(h_1(\tilde{X}_1,e_1,W_1)) = 0$, we have:

$$\begin{aligned} & Var[h_1(\tilde{X}_1,e_1,W_1)] \\ = \; & E[h_1(\tilde{X}_1,e_1,W_1)h_1(\tilde{X}_1,e_1,W_1)'] \\ = \; & E\left[\int_{\mathbb{R}^{qw}} e^{-it'W_1}e_1E[e^{it'W_2}\tilde{X}_2]d\mu(t)\int_{\mathbb{R}^{qw}} e^{-it'W_1}e_1E[e^{it'W_2}\tilde{X}_2']d\mu(t)\right] \end{aligned}$$

which is nonsingular since $\tilde{X}_2$ are not perfectly multicollinear (from Assumption 1(*iii*)). Following Theorem 7.1 in Hoeffding (1948a), the asymptotic distribution for U-statistics yields:

$$\sqrt{n}(B_n' - 0) \xrightarrow{d} \mathcal{N}(0, 4Var[h_1(\tilde{X}_1,e_1,W_1)])$$

Thus,

$$\sqrt{n}(\tilde{\beta}-\beta_0) \xrightarrow{d} N(0, [E(\kappa_{j,l}\tilde{X}_j\tilde{X}_l')]^{-1}Var[h_1(\tilde{X}_j,e_j,W_j)]E(\kappa_{j,l}\tilde{X}_j\tilde{X}_l')]^{-1})$$

(iii) We conclude the proof by showing that the above-mentioned WLLN and CLT for U-statistics apply under Assumption 2.

- From Theorem 7.1 in Hoeffding (1948a), a CLT applies to B_n' if: for any $1 \leq m \leq p$, $E[(\kappa_{1,2}\tilde{X}_{1,m}e_2 + \kappa_{2,1}\tilde{X}_{2,m}e_1)^2] < \infty$ and $Var[h_1(\tilde{X}_{1,m},e_1,W_1)]$ is positive definite. The second condition has already been shown in step (ii). We now show that the first condition holds under Assumption 2.

$$\begin{aligned} & E[(\kappa_{1,2}\tilde{X}_{1,m}e_2 + \kappa_{2,1}\tilde{X}_{2,m}e_1)^2] \\ =& E[(\kappa_{1,2}\tilde{X}_{1,m}e_2)^2] + E[(\kappa_{2,1}\tilde{X}_{2,m}e_1)^2] + 2E[\kappa_{1,2}\tilde{X}_{1,m}e_2\kappa_{2,1}\tilde{X}_{2,m}e_1] \\ =& I_1 + I_2 + 2I_3 \quad \text{with obvious notations.} \end{aligned}$$

Notice that $I_1 = I_2$. We now study I_1 and I_3 separately.

$$\begin{aligned}
I_1 &= E\left[\left(\int_{R^{q_w}} e^{it'(W_1-W_2)}d\mu(t)\right)^2 \tilde{X}_{1,m}^2 e_2^2\right] \\
&= \int_{R^{q_w}}\int_{R^{q_w}} E[e^{i(t+s)'(W_1-W_2)}\tilde{X}_{1,m}^2 e_2^2]d\mu(t)d\mu(s) \\
&= \int_{R^{q_w}}\int_{R^{q_w}} E[\tilde{X}_{1,m}^2 e^{i(t+s)'W_1} e_2^2 e^{-i(t+s)'W_2}]d\mu(t)d\mu(s) \\
&= \int_{R^{q_w}}\int_{R^{q_w}} E[\tilde{X}_{1,m}^2 e^{i(t+s)'W_1}]E[e_2^2 e^{-i(t+s)'W_2}]d\mu(t)d\mu(s) \\
&= \int_{R^{q_w}}\int_{R^{q_w}} E[E[\tilde{X}_{1,m}^2|W_1]e^{i(t+s)'W_1}]E[e_1^2 e^{-i(t+s)'W_1}]d\mu(t)d\mu(s) \\
&= \int_{R^{q_w}}\int_{R^{q_w}} \left(\int_{R^{q_w}} E[\tilde{X}_{1,m}^2|W_1]f_W(W_1)e^{i(t+s)'W_1}dW_1\right) E[e_1^2 e^{-i(t+s)'W_1}]d\mu(t)d\mu(s) \\
&= (2\pi)^{q_w}\int_{R^{q_w}}\int_{R^{q_w}} \mathcal{F}\{E[\tilde{X}_{1,m}^2|W_1]f_W(W_1)\}(t+s)E[e_1^2 e^{-i(t+s)'W_1}]d\mu(t)d\mu(s) \\
&= (2\pi)^{2q_w}\int_{R^{q_w}}\int_{R^{q_w}} \mathcal{F}\{E[\tilde{X}_{1,m}^2|W_1]f_W(W_1)\}(t+s)\mathcal{F}\{E[e_1^2|W_1]f_W(W_1)\}(-t-s)d\mu(t)d\mu(s)
\end{aligned}$$

with $\mathcal{F}$ the Fourier transform as defined in (1.1). Following Champeney (1987) p48, we need to ensure $E[\tilde{X}_{1,m}^2|W_1 = .]f_W(.)$ and $E[e_1^2|W_1 = .]f_W(.)$ are L_q for some $q \in [1,2]$ to ensure the existence of the corresponding Fourier transforms. These hold under Assumptions 2. Since $\mu(.)$ is a CDF, it follows that $I_1 < \infty$.

$$\begin{aligned}
I_3 &= E[\kappa_{1,2}\tilde{X}_{1,m}e_2\kappa_{2,1}\tilde{X}_{2,m}e_1] \\
&= E\left[\left(\int_{R^{q_w}} e^{it'(W_1-W_2)}d\mu(t)\right)\left(\int_{R^{q_w}} e^{is'(W_2-W_1)}d\mu(s)\right)\tilde{X}_{1,m}\tilde{X}_{2,m}e_1e_2\right] \\
&= \int_{R^{q_w}}\int_{R^{q_w}} E[e^{i(t-s)'(W_1-W_2)}\tilde{X}_{1,m}\tilde{X}_{2,m}e_1e_2]d\mu(t)d\mu(s) \\
&= \int_{R^{q_w}}\int_{R^{q_w}} E[\tilde{X}_{1,m}e_1e^{i(t-s)'W_1}\tilde{X}_{2,m}e_2e^{-i(t-s)'W_2}]d\mu(t)d\mu(s) \\
&= \int_{R^{q_w}}\int_{R^{q_w}} E[\tilde{X}_{1,m}e_1e^{i(t-s)'W_1}]E[\tilde{X}_{2,m}e_2e^{-i(t-s)'W_2}]d\mu(t)d\mu(s) \\
&= \int_{R^{q_w}}\int_{R^{q_w}} E[E[\tilde{X}_{1,m}e_1|W_1]e^{i(t-s)'W_1}]E[\tilde{X}_{1,m}e_1e^{-i(t-s)'W_1}]d\mu(t)d\mu(s) \\
&= \int_{R^{q_w}}\int_{R^{q_w}} \left(\int_{R^{q_w}} E[\tilde{X}_{1,m}e_1|W_1]f_W(W_1)e^{i(t-s)'W_1}dW_1\right) E[\tilde{X}_{1,m}e_1e^{-i(t-s)'W_1}]d\mu(t)d\mu(s) \\
&= (2\pi)^{2q_w}\int_{R^{q_w}}\int_{R^{q_w}} |\mathcal{F}\{E[\tilde{X}_{1,m}e_1|W_1]f_W(W_1)\}(t-s)|^2d\mu(t)d\mu(s)
\end{aligned}$$

Similarly, we need to ensure $E[\tilde{X}_{1,m}e_1|W_1 = .]f_W(.)$ is L_q for some $q \in [1,2]$, which holds under Assumptions 2. And it follows that $I_3 < \infty$.

- To justify that a WLLN for U-statistics applies to A_n, we follow section 1.3 in Bose and Chatterjee (2018) and show that a CLT for U-statistics applies by showing: for any

$1 \leq m_1, m_2 \leq p$, $E(\kappa_{1,2}\tilde{X}_{1,m_1}\tilde{X}_{2,m_2} + \kappa_{2,1}\tilde{X}_{2,m_1}\tilde{X}_{1,m_2}) < \infty$.

$$\begin{aligned}
& E[(\kappa_{1,2}\tilde{X}_{1,m_1}\tilde{X}_{2,m_2} + \kappa_{2,1}\tilde{X}_{2,m_1}\tilde{X}_{1,m_2})^2] \\
= \;& E[(\kappa_{1,2}\tilde{X}_{1,m_1}\tilde{X}_{2,m_2})^2] + E[(\kappa_{2,1}\tilde{X}_{2,m_1}\tilde{X}_{1,m_2})^2] + 2E[\kappa_{1,2}\tilde{X}_{1,m_1}\tilde{X}_{2,m_2}\kappa_{2,1}\tilde{X}_{2,m_1}\tilde{X}_{1,m_2}] \\
= \;& I'_1 + I'_2 + 2I'_3 \quad \text{with obvious notations.}
\end{aligned}$$

Notice that $I'_2 = I'_1$. We now study I'_1 and I'_3 separately.

$$\begin{aligned}
I'_1 &= E\left[\left(\int_{R^{q_w}} e^{it'(W_1-W_2)}d\mu(t)\right)^2 \tilde{X}^2_{1,m_1}\tilde{X}^2_{2,m_2}\right] \\
&= \int_{R^{q_w}}\int_{R^{q_w}} E[e^{i(t+s)'(W_1-W_2)}\tilde{X}^2_{1,m_1}\tilde{X}^2_{2,m_2}]d\mu(t)d\mu(s) \\
&= \int_{R^{q_w}}\int_{R^{q_w}} E[\tilde{X}^2_{1,m_1}e^{i(t+s)'W_1}\tilde{X}^2_{2,m_2}e^{-i(t+s)'W_2}]d\mu(t)d\mu(s) \\
&= \int_{R^{q_w}}\int_{R^{q_w}} E[\tilde{X}^2_{1,m_1}e^{i(t+s)'W_1}]E[\tilde{X}^2_{2,m_2}e^{-i(t+s)'W_2}]d\mu(t)d\mu(s) \\
&= \int_{R^{q_w}}\int_{R^{q_w}} E[E[\tilde{X}^2_{1,m_1}|W_1]e^{i(t+s)'W_1}]E[\tilde{X}^2_{1,m_2}e^{-i(t+s)'W_1}]d\mu(t)d\mu(s) \\
&= \int_{R^{q_w}}\int_{R^{q_w}} \left(\int_{R^{q_w}} E[\tilde{X}^2_{1,m_1}|W_1]f_W(W_1)e^{i(t+s)'W_1}dW_1\right) E[\tilde{X}^2_{1,m_2}e^{-i(t+s)'W_1}]d\mu(t)d\mu(s) \\
&= (2\pi)^{q_w}\int_{R^{q_w}}\int_{R^{q_w}} \mathcal{F}\{E[\tilde{X}^2_{1,m_1}|W_1]f_W(W_1)\}(t+s)E[\tilde{X}^2_{1,m_2}e^{-i(t+s)'W_1}]d\mu(t)d\mu(s) \\
&= (2\pi)^{2q_w}\int_{R^{q_w}}\int_{R^{q_w}} \mathcal{F}\{E[\tilde{X}^2_{1,m_1}|W_1]f_W(W_1)\}(t+s)\mathcal{F}\{E[\tilde{X}^2_{1,m_2}|W_1]f_W(W_1)\}(-t-s)d\mu(t)d\mu(s)
\end{aligned}$$

Following Champeney (1987) once again, we need to ensure $E[\tilde{X}^2_{1,m_1}|W_1 = .]f_W(.)$ and $E[\tilde{X}^2_{1,m_2}|W_1 = .]f_W(.)$ are L_q for some $q \in [1,2]$. These hold under Assumptions 2. And it follows that $I'_1 < \infty$.

$$\begin{aligned}
I'_3 &= E[\kappa_{1,2}\tilde{X}_{1,m_1}\tilde{X}_{2,m_2}\kappa_{2,1}\tilde{X}_{2,m_1}\tilde{X}_{1,m_2}] \\
&= E\left[\left(\int_{R^{q_w}} e^{it'(W_1-W_2)}d\mu(t)\right)\left(\int_{R^{q_w}} e^{is'(W_2-W_1)}d\mu(s)\right)\tilde{X}_{1,m_1}\tilde{X}_{2,m_2}\tilde{X}_{2,m_1}\tilde{X}_{1,m_2}\right] \\
&= \int_{R^{q_w}}\int_{R^{q_w}} E[e^{i(t-s)'(W_1-W_2)}\tilde{X}_{1,m_1}\tilde{X}_{2,m_2}\tilde{X}_{2,m_1}\tilde{X}_{1,m_2}]d\mu(t)d\mu(s) \\
&= \int_{R^{q_w}}\int_{R^{q_w}} E[\tilde{X}_{1,m_1}\tilde{X}_{1,m_2}e^{i(t-s)'W_1}\tilde{X}_{2,m_2}\tilde{X}_{2,m_1}e^{-i(t-s)'W_2}]d\mu(t)d\mu(s) \\
&= \int_{R^{q_w}}\int_{R^{q_w}} E[\tilde{X}_{1,m_1}\tilde{X}_{1,m_2}e^{i(t-s)'W_1}]E[\tilde{X}_{2,m_2}\tilde{X}_{2,m_1}e^{-i(t-s)'W_2}]d\mu(t)d\mu(s) \\
&= \int_{R^{q_w}}\int_{R^{q_w}} E[E[\tilde{X}_{1,m_1}\tilde{X}_{1,m_2}|W_1]e^{i(t-s)'W_1}]E[\tilde{X}_{1,m_2}\tilde{X}_{1,m_1}e^{-i(t-s)'W_1}]d\mu(t)d\mu(s) \\
&= \int_{R^{q_w}}\int_{R^{q_w}} \left(\int_{R^{q_w}} E[\tilde{X}_{1,m_1}\tilde{X}_{1,m_2}|W_1]f_W(W_1)e^{i(t-s)'W_1}dW_1\right) E[\tilde{X}_{1,m_2}\tilde{X}_{1,m_1}e^{-i(t-s)'W_1}]d\mu(t)d\mu(s) \\
&= (2\pi)^{2q_w}\int_{R^{q_w}}\int_{R^{q_w}} |\mathcal{F}\{E[\tilde{X}_{1,m_1}\tilde{X}_{1,m_2}|W_1]f_W(W_1)\}(t-s)|^2d\mu(t)d\mu(s)
\end{aligned}$$

We need to ensure $E[\tilde{X}_{1,m_1}\tilde{X}_{1,m_2}|W_1 = .]f_W(.)$ is L_q for some $q \in [1,2]$, which holds under Assumptions 2. And it follows that $I_3' < \infty$. □

1.5 Proof of Theorem 3.1

Proof. Recall that

$$\hat{\tilde{y}}_i = y_i - \hat{g}_y(Z_i) = y_i - E(y_i|Z_i) + E(y_j|Z_i) - \hat{g}_y(Z_i)$$

For the first two terms of the right hand side, we have

$$\begin{array}{rcl} y_i - E(y_i|Z_i) & = & (X_i - E(X_i|Z_i))'\beta_0 + e_i \\ & = & (X_i - \hat{g}_X(Z_i))'\beta_0 + (\hat{g}_X(Z_i) - E(X_i|Z_i))'\beta_0 + e_i \end{array}$$

In matrix form, the feasible estimator writes:

$$\hat{\beta}_n = [\hat{\tilde{X}}'\tilde{\kappa}\hat{\tilde{X}}]^{-1}\hat{\tilde{X}}'\tilde{\kappa}\hat{\tilde{y}} = \left[\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\hat{\tilde{X}}_j\widehat{\tilde{X}_l}'\right]^{-1}\left[\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\hat{\tilde{X}}_j\hat{\tilde{y}}_l\right]$$

Define $C_n = \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\hat{\tilde{X}}_j\widehat{\tilde{X}_l}'$. We get:

$$\begin{array}{rcl} \hat{\beta}_n & = & [C_n]^{-1}[\dfrac{1}{n(n-1)}\displaystyle\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\hat{\tilde{X}}_j[y_l - E(y_l|Z_l) + E(y_l|Z_l) - \hat{g}_y(Z_l)]] \\ & = & [C_n]^{-1}[\dfrac{1}{n(n-1)}\displaystyle\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\hat{\tilde{X}}_j[(X_l - \hat{g}_X(Z_l))'\beta_0 \\ & & +(\hat{g}_X(Z_l) - E(X_l|Z_l))'\beta_0 + e_l + E(y_l|Z_l) - \hat{g}_y(Z_l)]] \\ & = & [C_n]^{-1}[\dfrac{1}{n(n-1)}\displaystyle\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\hat{\tilde{X}}_j[\hat{\tilde{X}}_l'\beta_0 \\ & & +(\hat{g}_X(Z_l) - E(X_l|Z_l))'\beta_0 + e_l + E(y_l|Z_l) - \hat{g}_y(Z_l)]] \\ & = & \beta_0 + [C_n]^{-1}[\dfrac{1}{n(n-1)}\displaystyle\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\hat{\tilde{X}}_j[(\hat{g}_X(Z_l) - E(X_l|Z_l))'\beta_0 \\ & & +e_l + E(y_l|Z_l) - \hat{g}_y(Z_l)]] \end{array}$$

Consider now,

$$\begin{aligned}
A_n - C_n &= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{X}_j\tilde{X}_l' \\
&\quad -\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}[\tilde{X}_j + O_p(v_n)][\tilde{X}_l + O_p(v_n)]' \\
&= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{X}_j\tilde{X}_l' \\
&\quad -\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}[\tilde{X}_j\tilde{X}_l' + O_p(v_n)\tilde{X}_l' + \tilde{X}_j O_p(v_n) + O_p(v_n^2)] \\
&= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}[O_p(v_n)\tilde{X}_l' + \tilde{X}_j O_p(v_n) + O_p(v_n^2)] \\
&\overset{P}{\to} 0
\end{aligned}$$

Hence, we have $\texttt{Plim}C_n = A$, since we showed in the proof of Proposition 2 that $\texttt{Plim}A_n = A$.

Define now the following quantities:

$$\begin{aligned}
D_n &= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\hat{\tilde{X}}_j[(\hat{g}_X(Z_l) - E(X_l|Z_l))'\beta_0 + e_l + E(y_l|Z_l) - \hat{g}_y(Z_l)] \\
E_n &= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\hat{\tilde{X}}_j(\hat{g}_X(Z_l) - E(X_l|Z_l))'\beta_0 \\
F_n &= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\hat{\tilde{X}}_j e_l \\
G_n &= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\hat{\tilde{X}}_j[E(y_l|Z_l) - \hat{g}_y(Z_l)]
\end{aligned}$$

We have, $D_n = E_n + F_n + G_n$. For consistency, we show that the probability limits for E_n, F_n, and G_n are all zero.

$$
\begin{aligned}
E_n &= \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{l \neq j}^{n} \kappa_{j,l} \hat{\tilde{X}}_j (\hat{g}_X(Z_l) - E(X_l|Z_l))' \beta_0 \\
&= \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{l \neq j}^{n} \kappa_{j,l} [\tilde{X}_j + Op(v_n)] (Op(v_n))' \beta_0 \\
&\xrightarrow{P} 0 \\
F_n &= \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{l \neq j}^{n} \kappa_{j,l} [\tilde{X}_j + Op(v_n)] e_l \\
&= \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{l \neq j}^{n} \kappa_{j,l} [\tilde{X}_j e_l + Op(v_n) e_l] \\
&\xrightarrow{P} 0
\end{aligned}
$$

since, from the proof of Proposition 2, $E(\kappa_{j,l} \tilde{X}_i e_l) = 0$.

$$
\begin{aligned}
G_n &= \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{l \neq j}^{n} \kappa_{j,l} \hat{\tilde{X}}_j [E(y_l|Z_l) - \hat{g}_y(Z_l)] \\
&= \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{l \neq j}^{n} \kappa_{j,l} [\tilde{X}_j + Op(v_n)] [Op(v_n)] \\
&\xrightarrow{P} 0
\end{aligned}
$$

All in all, we have $\texttt{Plim} D_n = \texttt{Plim}(E_n + F_n + G_n) = 0$, so $\hat{\beta}_n \xrightarrow{P} \beta_0$.

In addition, we have:

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = [C_n]^{-1} \sqrt{n} [E_n + F_n + G_n]$$

And we study each term separately:

$$\begin{aligned}
\sqrt{n}F_n &= \sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}[\tilde{X}_j + O_p(v_n)]e_l \\
&= \sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}[\kappa_{j,l}\tilde{X}_j e_l + \kappa_{j,l}O_p(v_n)e_l] \\
&= \sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{X}_j e_l + \sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}O_p(v_n)e_l \\
&= \sqrt{n}B_n + \sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}O_p(v_n)e_l \\
\sqrt{n}E_n &= \sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\hat{\tilde{X}}_j(\hat{g}_X(Z_l) - E(X_l|Z_l))'\beta_0 \\
&= \sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}[\tilde{X}_j + Op(v_n)](Op(v_n))'\beta_0
\end{aligned}$$

$$\begin{aligned}
\sqrt{n}G_n &= \sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\hat{\tilde{X}}_j[E(y_l|Z_l) - \hat{g}_y(Z_l)] \\
&= \sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}[\tilde{X}_j + O_p(v_n)][O_p(v_n)]
\end{aligned}$$

Thus, if we can show that

$$\sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{X}_j O_p(v_n) = o_p(1)$$

then it will follow that $\sqrt{n}E_n$ and $\sqrt{n}G_n$ are $o_p(1)$. Indeed,

$$\sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{X}_j O_p(v_n) = [\sqrt{n}\frac{1}{n}\sum_{j}^{n}\tilde{X}_j\frac{1}{n-1}\sum_{l\neq j}^{n}\kappa_{j,l}]O_p(v_n)$$

For each j,

$$|\frac{1}{n-1}\sum_{l\neq j}^{n}\kappa_{j,l}| \leq \frac{1}{n-1}\sum_{l\neq j}^{n}|\kappa_{j,l}|$$

From the definition for $\kappa_{j,l}$, $|\kappa_{j,l}| = |\int_{R^{qw}} e^{it'(W_j-W_l)}d\mu(t)|$. From the properties of Lebesgue integral, we have:

$$|\int_{R^{qw}} e^{it'(W_j-W_l)}d\mu(t)| < \int_{R^{qw}} |e^{it'(W_j-W_l)}|d\mu(t)$$

Since $|e^{it'(W_j-W_l)}| = |cos(t'(W_j - W_l)) + isin(t'(W_j - W_l))|$ and Assumption $1(vi)$, we have

$$|e^{it'(W_j-W_l)}| = |cos(t'(W_j - W_l))| \leq 1$$

Thus,

$$\begin{aligned} |\kappa_{j,l}| \leq \int_{R^{qw}} d\mu(t) = M \quad w.p.1 \quad &\Rightarrow \quad |\frac{1}{n-1}\sum_{l\neq j}^{n} \kappa_{j,l}| \leq M \quad w.p.1 \\ &\Rightarrow \quad \frac{1}{n-1}\sum_{l\neq j}^{n} \kappa_{j,l} = O_p(1) \end{aligned}$$

Hence,

$$\sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n} \kappa_{j,l}\tilde{X}_j O_p(v_n) = [\sqrt{n}\frac{1}{n}\sum_{j}^{n} \tilde{X}_j O_p(1)]O_p(v_n)$$

Also, $\sqrt{n}\frac{1}{n}\sum_{j}^{n}\tilde{X}_j = O_p(1)$. We have $\sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{X}_j O_p(v_n) = O_p(v_n)$. Thus,

$$\sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n} \kappa_{j,l}\tilde{X}_j O_p(v_n) \qquad \text{and} \qquad \sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n} \kappa_{j,l} O_p(v_n) e_l$$

are $o_p(1)$. Hence, we have:

$$\sqrt{n}(\hat{\beta} - \beta_0) = [A_n + o_p(1)]^{-1}\sqrt{n}[B_n + o_p(1)]$$

and we conclude that the feasible estimator has the same asymptotic distribution as the infeasible one. □

1.6 Consistent estimator of the asymptotic variance (3.10)

$$\begin{aligned} & Var[\int_{R^{qw}} e^{-it'W_j} e_j(\beta_0) E[e^{it'W_l}\tilde{X}_l] d\mu(t)] \\ = \;& Var[E_l(\int_{R^{qw}} e^{it'(W_l - W_j)} e_j(\beta_0)\tilde{X}_l d\mu(t))] \\ = \;& Var[E_l(k(W_l - W_j)\tilde{X}_l e_j(\beta_0))] \\ = \;& E[(E_l[k(W_l - W_j)\tilde{X}_l])(E_l[k(W_l - W_j)\tilde{X}_l'])[e_j(\beta_0)]^2] \end{aligned}$$

Under heteroskedasticity, the estimator for variance of the feasible-RSMD is

$$\begin{aligned} & [\sum_{j=1}^{n}\sum_{l\neq j}^{n} \kappa_{j,l}\hat{\tilde{X}}_j \widehat{\tilde{X}_l}']^{-1} \sum_{j=1}^{n}((\sum_{l=1}^{n}\kappa_{j,l}\widehat{\tilde{X}_l})(\sum_{l=1}^{n}\kappa_{j,l}\widehat{\tilde{X}_l}')\hat{e}_j^2)[\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\hat{\tilde{X}}_j\widehat{\tilde{X}_l}']^{-1} \\ = \;& [\hat{\tilde{X}}'\tilde{\kappa}\hat{\tilde{X}}]^{-1}\hat{\tilde{X}}'\tilde{\kappa}\Omega_n\tilde{\kappa}\hat{\tilde{X}}[\hat{\tilde{X}}'\tilde{\kappa}\hat{\tilde{X}}]^{-1} \end{aligned}$$

with Ω_n the conventional variance matrix of residuals from $\widehat{\tilde{y}}_i - \hat{\tilde{X}}_i'\hat{\beta}_n$. The following proves that the two expressions are the same.

Proof. $\hat{\tilde{X}}'\tilde{\kappa}\Omega_n\tilde{\kappa}\hat{\tilde{X}} = \hat{\tilde{X}}'\tilde{\kappa}\begin{bmatrix}\sigma_1^2 & ... & 0\\ 0 & ... & \sigma_n^2\end{bmatrix}\tilde{\kappa}\hat{\tilde{X}} = \begin{bmatrix}\hat{\tilde{X}}_1 & ... & \hat{\tilde{X}}_n\end{bmatrix}\begin{bmatrix}k_{11} & k_{12} & ... & k_{1n}\\ k_{21} & k_{22} & ... & k_{2n}\\ ... & ... & ... & ...\\ k_{n1} & k_{n2} & ... & k_{nn}\end{bmatrix}\begin{bmatrix}\sigma_1^2 & ... & 0\\ 0 & ... & \sigma_n^2\end{bmatrix}\tilde{\kappa}\hat{\tilde{X}}$

$= \begin{bmatrix}\sum_{l=1}^n \hat{\tilde{X}}_l k_{l1} & ... & \sum_{l=1}^n \hat{\tilde{X}}_l k_{ln}\end{bmatrix}\begin{bmatrix}\sigma_1^2 & ... & 0\\ 0 & ... & \sigma_n^2\end{bmatrix}\tilde{\kappa}\hat{\tilde{X}}$

$= \begin{bmatrix}\sum_{l=1}^n \hat{\tilde{X}}_l k_{l1}\sigma_1^2 & ... & \sum_{l=1}^n \hat{\tilde{X}}_l k_{ln}\sigma_n^2\end{bmatrix}\tilde{\kappa}\hat{\tilde{X}}$

$= \begin{bmatrix}\sum_{l=1}^n \hat{\tilde{X}}_l k_{l1}\sigma_1^2 & ... & \sum_{l=1}^n \hat{\tilde{X}}_l k_{ln}\sigma_n^2\end{bmatrix}\begin{bmatrix}\sum_{l=1}^n \hat{\tilde{X}}'_l k_{l1}\sigma_1^2\\ ...\\ \sum_{l=1}^n \hat{\tilde{X}}'_l k_{ln}\sigma_n^2\end{bmatrix}$

$= \begin{bmatrix}\sum_{l=1}^n \hat{\tilde{X}}_l k_{l1}\sigma_1^2 & ... & \sum_{l=1}^n \hat{\tilde{X}}_l k_{ln}\sigma_n^2\end{bmatrix}\begin{bmatrix}\sum_{l=1}^n \hat{\tilde{X}}'_l k_{l1}\\ ...\\ \sum_{l=1}^n \hat{\tilde{X}}'_l k_{ln}\end{bmatrix}$

$= \sum_{j=1}^n(\sum_{l=1}^n \hat{\tilde{X}}_l k_{lj})(\sum_{l=1}^n \hat{\tilde{X}}'_l k_{lj})\sigma_j^2$

□

• Consistent estimator under homoskedasticity:
When the error term is homoskedastic, the estimator for variance of feasible-RSMD is

$$\begin{aligned} & \hat{\sigma}^2[\sum_{j=1}^n\sum_{l\neq j}^n \kappa_{j,l}\hat{\tilde{X}}_j\widehat{\tilde{X}_l}']^{-1}\sum_{j=1}^n((\sum_{l=1}^n \kappa_{j,l}\widehat{\tilde{X}_l})(\sum_{l=1}^n \kappa_{j,l}\widehat{\tilde{X}_l}'))[\sum_{j=1}^n\sum_{l\neq j}^n \kappa_{j,l}\hat{\tilde{X}}_j\widehat{\tilde{X}_l}']^{-1} \\ = \quad & \hat{\sigma}^2[\hat{\tilde{X}}'\tilde{\kappa}\hat{\tilde{X}}]^{-1}\hat{\tilde{X}}'\tilde{\kappa}\tilde{\kappa}\hat{\tilde{X}}[\hat{\tilde{X}}'\tilde{\kappa}\hat{\tilde{X}}]^{-1} \end{aligned}$$

with $\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^n \sigma_j^2$.

1.7 Choosing standard normal distribution

$\kappa_{j,l} = k(W_j - W_l) = \int_{R^{q_w}} e^{it'(W_j - W_l)}d\mu(t) = e^{-u^2/2}$ with $u = (W_j - W_l)$, when $\mu(t)$ is a standard normal distribution on t and W is of dimension 1.

Proof.

$$\begin{aligned}
\kappa_{j,l} &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}^{q_w}} e^{-t^2/2} e^{itu} dt \\
&= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}^{q_w}} e^{-t^2/2+itu} dt \\
&= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}^{q_w}} e^{-1/2(t^2-2itu+u^2i^2-u^2i^2)} dt \\
&= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}^{q_w}} e^{-1/2(t^2-2itu+u^2i^2)+1/2u^2i^2} dt \\
&= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}^{q_w}} e^{-1/2(t-iu)^2+u^2i^2/2} dt \\
&= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}^{q_w}} e^{-1/2(t-iu)^2} dt e^{-u^2/2} \\
&= \frac{1}{\sqrt{2\pi}} \sqrt{2\pi} e^{-u^2/2} \\
&= e^{-u^2/2}
\end{aligned}$$

With $\int_{\mathbb{R}^{q_w}} e^{-1/2(t-iu)^2} dt = \sqrt{2\pi}$, the integral of Gaussian function. □

2 Two-step R-SMD estimator

We start this section by introducing the (infeasible) efficient SMD estimator and presenting its asymptotic properties. We then introduce the two-step R-SMD estimator and derive its asymptotic properties.

2.1 The infeasible efficient SMD estimator

Let us introduce the efficient weighting matrix, $Var[e_i(\beta_0)|W_i]f(W_i)$, and its nonparametric estimator

$$\hat{\omega}_n(W_i, \hat{\beta}_1) \equiv \frac{1}{nb^{q_w}} \sum_{k=1}^{n} [\hat{e}_k(\hat{\beta}_1)]^2 L\left(\frac{W_i - W_k}{b}\right)$$

where $\hat{\beta}_1$ denotes a (first-step) consistent estimator such as $\tilde{\beta}_n$ or $\hat{\beta}_n$ and $L(.)$ a second-order product kernel with b a vanishing bandwidth as defined in Assumption 6 below.

The infeasible efficient estimator $\tilde{\beta}_e$ is defined as the minimizer of $M_{n,\tilde{h},b}(\beta)$,

$$M_{n,\tilde{h},b} \equiv \frac{1}{n(n-1)} \sum_{l=1} \sum_{j\neq l} \hat{\omega}_n^{-1/2}(W_j, \hat{\beta}_1)\hat{\omega}_n^{-1/2}(W_l, \hat{\beta}_1)[\tilde{y}_j - \tilde{X}_j'\beta][\tilde{y}_l - \tilde{X}_l'\beta]\tilde{h}^{-q_w} k\left(\frac{W_j - W_l}{\tilde{h}}\right)$$

and is obtained in closed-form by solving the FOC as:

$$\begin{aligned}\tilde{\beta}_{e,n} &\equiv [\sum_{l=1}\sum_{j\neq l}\hat{\omega}_n^{-1/2}(W_j,\hat{\beta}_1)\hat{\omega}_n^{-1/2}(W_l,\hat{\beta}_1)\kappa_{j,l,\tilde{h}}\tilde{X}_j\tilde{X}_l']^{-1} \\ &\quad\times[\sum_{l=1}\sum_{j\neq l}\hat{\omega}_n^{-1/2}(W_j,\hat{\beta}_1)\hat{\omega}_n^{-1/2}(W_l,\hat{\beta}_1)\kappa_{j,l,\tilde{h}}\tilde{X}_j\tilde{y}_l] \qquad (2.2)\end{aligned}$$

with $\kappa_{j,l,\tilde{h}} \equiv \tilde{h}^{-q_w}k\left(\frac{W_j-W_l}{\tilde{h}}\right)$.

In order to derive the asymptotic properties of $\tilde{\beta}_{e,n}$ we need additional regularity assumptions.

Assumption 6. *(Vanishing bandwidth and Regularity of the kernel $L(.)$)*
(i) The bandwidth $\tilde{h} > 0$ is $o(1)$ with $\sqrt{n}(\tilde{h}^4 + [1/(n\tilde{h}^{q_w})]) = o(1)$.
(ii) $L(.)$ is a product kernel based on a second-order univariate kernel $l(.)$ such that

$$\begin{aligned} L\left(\frac{W_i - W_j}{b}\right) &= \Pi_{s=1}^{q_w} l\left(\frac{W_{s,i}-W_{s,j}}{b_s}\right) \\ \textit{with} \qquad b \equiv \Pi_{s=1}^{q_w} b_s \quad \textit{and} \quad & \sqrt{n}\left(\sum_{s=1}^{q_w} b_s^4 + \left[\frac{1}{nb_1 \ldots b_{q_w}}\right]\right) = o(1)\end{aligned}$$

Proposition 5. *(Consistency and Asymptotic normality of $\tilde{\beta}_{e,n}$)*
Under Assumptions 1 to 3 and 6, $\tilde{\beta}_{e,n}$ is consistent for β_0, that is $\tilde{\beta}_{e,n} \xrightarrow{p} \beta_0$, asymptotically normally distributed, and efficient as its asymptotic variance reaches the semi-parametric efficiency bound,

$$\sqrt{n}(\tilde{\beta}_{e,n} - \beta_0) \xrightarrow{d} \mathcal{N}(0, [E[Var[e_j(\beta_0)|W_j]^{-1}E(\tilde{X}_j|W_j)E(\tilde{X}_j'|W_j)]]^{-1})$$

2.2 The two-step R-SMD estimator

Consider now the feasible counterpart of $\tilde{\beta}_{e,n}$ obtained as the minimizer of

$$\frac{1}{n(n-1)}\sum_{l=1}\sum_{j\neq l}\hat{\omega}_n^{-1/2}(W_j,\hat{\beta}_1)\hat{\omega}_n^{-1/2}(W_l,\hat{\beta}_1)[\widehat{\tilde{y}}_j - \widehat{\tilde{X}}_j'\beta][\widehat{\tilde{y}}_l - \widehat{\tilde{X}}_l'\beta]\tilde{h}^{-q_w}k\left(\frac{W_j-W_l}{\tilde{h}}\right)$$

It is denoted $\hat{\beta}_{e,n}$ and obtained in closed-form as:

$$\begin{aligned}\hat{\beta}_{e,n} &= [\sum_{j=1}^{n}\sum_{l\neq j}^{n}\omega_n^{-1/2}(W_j,\hat{\beta}_1)\hat{\omega}_n^{-1/2}(W_l,\hat{\beta}_1)\kappa_{j,l,\tilde{h}}\widehat{\tilde{X}}_j\widehat{\tilde{X}}_l']^{-1} \\ &\quad\times[\sum_{j=1}^{n}\sum_{l\neq j}^{n}\omega_n^{-1/2}(W_j,\hat{\beta}_1)\hat{\omega}_n^{-1/2}(W_l,\hat{\beta}_1)\kappa_{j,l,\tilde{h}}\widehat{\tilde{X}}_j\widehat{\tilde{y}}_l] \qquad (2.3)\end{aligned}$$

Theorem 2.5. *(Asymptotic properties of $\hat{\beta}_{e,n}$)*
Under Assumptions 1 to 3 and 6, $\hat{\beta}_{e,n}$ is consistent for β_0, that is $\hat{\beta}_{e,n} \xrightarrow{p} \beta_0$ and

$$\sqrt{n}(\hat{\beta}_{e,n} - \beta_0 + H_n) \xrightarrow{d} \mathcal{N}(0, [E[Var[e_j(\beta_0)|W_j]^{-1}E(\tilde{X}_j|W_j)E(\tilde{X}_j'|W_j)]]^{-1})$$

where $H_n = \mathcal{O}_p(v_n)$.

2.3 Proofs of the theoretical results

Proof of Proposition 5

- Preliminary result #1:

Under the assumptions of Proposition 5, we have:

$$\begin{aligned}\hat{\omega}_n(W_i, \hat{\beta}_1) &= Var[e_i(\beta_0)|W_i]f(W_i) + \mathcal{O}_p\left(v_{b,n}\right) + o_p(1) \\ \text{with} \quad v_{b,n} &\equiv \sum_{s=1}^{q_w} b_s^2 + \left[\frac{1}{nb_1...b_{q_w}}\right]^{0.5}\end{aligned}$$

Proof. Recall that

$$\hat{g}_y(Z_i) - E(y_i|Z_i) = \mathcal{O}_p\left(v_n\right) \quad \text{and} \quad \hat{g}_X(Z_i) - E(X_i|Z_i) = \mathcal{O}_p\left(v_n\right)$$

with $v_n \equiv \sum_{s=1}^{q_z} h_s^2 + \left[\frac{1}{nh_1...h_{q_z}}\right]^{0.5}$. Hence, with $\hat{\beta}_1$ a consistent estimator of β_0, we have:

$$\begin{aligned}\hat{e}_k(\hat{\beta}_1) &= y_k - E(y_k|Z_k) - (X_k - E(X_k|Z_k))'\hat{\beta}_1 + E(y_k|Z_k) - \hat{g}_y(Z_k) \\ &\quad +(\hat{g}_X(Z_k) - E(X_k|Z_k))'\hat{\beta}_1 \\ &= y_k - E(y_k|Z_k) - (X_k - E(X_k|Z_k))'(\beta_0 + o_p(1)) + E(y_k|Z_k) - \hat{g}_y(Z_k) \\ &\quad +(\hat{g}_X(Z_k) - E(X_k|Z_k))'\hat{\beta}_1 \\ &= e_k(\beta_0) + E(y_k|Z_k) - \hat{g}_y(Z_k) + (\hat{g}_X(Z_k) - E(X_k|Z_k))'\hat{\beta}_1 + (X_k - E(X_k|Z_k))'o_p(1) \\ &= e_k(\beta_0) + \mathcal{O}_p\left(v_n\right) + \mathcal{O}_p\left(v_n\right)'\hat{\beta}_1 + (X_k - E(X_k|Z_k))'o_p(1) \\ &= e_k(\beta_0) + \mathcal{O}_p\left(v_n\right) + \mathcal{O}_p\left(v_n\right)'\left(\beta_0 + o_p(1)\right) + (X_k - E(X_k|Z_k))'o_p(1) \\ &= e_k(\beta_0) + \mathcal{O}_p\left(v_n\right) + (X_k - E(X_k|Z_k))'o_p(1)\end{aligned}$$

because $(X_k - E(X_k|Z_k) < \infty$ for all k, $X_k - E(X_k|Z_k) < \max_k\ (X_k - E(X_k|Z_k) = M$.
We then have:

$$\begin{aligned}\hat{\omega}_n(W_i, \hat{\beta}_1) &= \frac{1}{nb^{q_w}}\sum_{k=1}^{n}[e_k(\beta_0) + \mathcal{O}_p\left(v_n\right) + op(1)]^2 L\left(\frac{W_i - W_k}{b}\right) \\ &= \frac{1}{nb^{q_w}}\sum_{k=1}^{n}[e_k(\beta_0) + \mathcal{O}_p\left(v_n\right)]^2 L\left(\frac{W_i - W_k}{b}\right) \\ &= \frac{1}{nb^{q_w}}\sum_{k=1}^{n}[e_k(\beta_0)^2 + 2e_k(\beta_0)\mathcal{O}_p\left(v_n\right) + \mathcal{O}_p\left(v_n\right)^2]L\left(\frac{W_i - W_k}{b}\right) \\ &= \frac{1}{nb^{q_w}}\sum_{k=1}^{n}\{[e_k(\beta_0)^2]L\left(\frac{W_i - W_k}{b}\right) + [2e_k(\beta_0)\mathcal{O}_p\left(v_n\right) + \mathcal{O}_p\left(v_n\right)^2]L\left(\frac{W_i - W_k}{b}\right)\} \\ &= \frac{1}{nb^{q_w}}\sum_{k=1}^{n}[e_k(\beta_0)^2]L\left(\frac{W_i - W_k}{b}\right) + o_p(1)\end{aligned}$$

The first term of the RHS is the nonparametric estimator of $Var[e_i(\beta_0)|W_i]f(W_i)$ and from Li and Racine page 63,

$$\frac{1}{nb^{q_w}}\sum_{k=1}^{n}[e_k(\beta_0)^2]L\left(\frac{W_i-W_k}{b}\right)=Var[e_i(\beta_0)|W_i]f(W_i)+\mathcal{O}_p\left(v_{b,n}\right).$$

It then follows that:

$$\hat{\omega}_n(W_i,\hat{\beta}_1)=Var[e_i(\beta_0)|W_i]f(W_i)+\mathcal{O}_p\left(v_{b,n}\right)+o_p(1)\,.$$

□

• Preliminary result #2:
Let $\hat{\omega}_{n,j,1}\equiv\hat{\omega}_n(W_j,\hat{\beta}_1)$. Under the assumptions of Proposition 5, we have:

$$\begin{aligned}A_n &\equiv \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\tilde{X}_j\tilde{X}_l'\\ &\overset{P}{\to} A\equiv E[Var(e_j(\beta_0)|W_j)^{-1}E(\tilde{X}_j|W_j)E(\tilde{X}_j'|W_j)]\end{aligned}$$

Proof.

$$\begin{aligned}A_n &= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\tilde{X}_j\tilde{X}_l'\\ &= \frac{1}{n}\sum_{j=1}^{n}\hat{\omega}_{n,j,1}^{-1/2}\tilde{X}_j\left[\frac{1}{n-1}\sum_{l\neq j}^{n}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\tilde{X}_l'\right]\end{aligned}$$

Hence, A_n is a U-statistic. When the bandwidths $\tilde{h}$ and b converge to 0, the term in between the square brackets corresponds to the leave-one-out non-parametric estimator for $E(\omega_j^{-1/2}\tilde{X}_j'|W_j)f(W_j)$ with $\omega_j\equiv Var[e_j(\beta_0)|W_j]f(W_j)$. It follows that:

$$\begin{aligned}A_n &= \frac{1}{n}\sum_{j=1}^{n}[\omega_j+\mathcal{O}_p\left(v_{b,n}\right)+o_p(1)]^{-1/2}\tilde{X}_j\{\frac{1}{n-1}\sum_{l\neq j}^{n}[\omega_l+\mathcal{O}_p\left(v_{b,n}\right)+o_p(1)]^{-1/2}\kappa_{j,l,\tilde{h}}\tilde{X}_l'\}\\ &= \frac{1}{n}\sum_{j=1}^{n}[\omega_j+\mathcal{O}_p\left(v_{b,n}\right)+o_p(1)]^{-1/2}\tilde{X}_j[E(\omega_j^{-1/2}\tilde{X}_j'|W_j)f(W_j)+\mathcal{O}_p\left(v_{\tilde{h},n}\right)]\end{aligned}$$

with

$$v_{\tilde{h},n}\equiv\tilde{h}_s^2+\left[\frac{1}{n\tilde{h}^{q_w}}\right]^{0.5}.$$

Under the iid assumption maintained in Assumption 2, a LLN for U-statistics applies and the expected result follows:

$$\begin{aligned}
Plim\ A_n &= E[\omega_j^{-1/2}\tilde{X}_j E(\omega_j^{-1/2}\tilde{X}_j'|W_j)f(W_j)] \\
&= E[\omega_j^{-1}\tilde{X}_j E(\tilde{X}_j'|W_j)f(W_j)] \\
&= E[Var[e_j(\beta_0)|W_j]^{-1}f(W_j)^{-1}\tilde{X}_j E(\tilde{X}_j'|W_j)f(W_j)] \\
&= E[Var[e_j(\beta_0)|W_j]^{-1}\tilde{X}_j E(\tilde{X}_j'|W_j)] \\
&= E[Var[e_j(\beta_0)|W_j]^{-1}E(\tilde{X}_j|W_j)E(\tilde{X}_j'|W_j)]
\end{aligned}$$

□

• We now return to the proof of Theorem 5.

Proof. From the definition of $\tilde{\beta}_{e,n}$ and simple algebra, we have:

$$\begin{aligned}
\tilde{\beta}_{e,n} &= [A_n]^{-1}[\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\tilde{X}_j(\tilde{X}_l'\beta_0+e_l)] \\
&= [A_n]^{-1}[\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\tilde{X}_j\tilde{X}_l'\beta_0+\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\tilde{X}_je_l] \\
&= [A_n]^{-1}[\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\tilde{X}_j\tilde{X}_l'\beta_0] \\
&\quad +[A_n]^{-1}[\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\tilde{X}_je_l] \\
&= \beta_0+[A_n]^{-1}B_n \quad \text{with} \quad B_n\equiv\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\tilde{X}_je_l\,. \qquad (2.4)
\end{aligned}$$

From Preliminary result #2, we know that $A_n \xrightarrow{P} A$. Under Assumption 1(iv), A is nonsingular. To show that B_n is a U-statistic, notice that:

$$B_n=\frac{1}{n(n-1)}\sum_{j<l}^{n}(\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\tilde{X}_je_l+\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{l,j,\tilde{h}}\tilde{X}_le_j)$$

with $\kappa_{j,l,\tilde{h}}\equiv\tilde{h}^{-q_w}k\left(\frac{W_j-W_l}{\tilde{h}}\right)$. Define now

$$h(\tilde{\omega}_1,\tilde{x}_1,\epsilon_1,w_1;\tilde{\omega}_2,\tilde{x}_2,\epsilon_2,w_2)=\tilde{\omega}_1^{-1/2}\tilde{\omega}_2^{-1/2}\kappa_{1,2,\tilde{h}}\tilde{x}_1\epsilon_2+\tilde{\omega}_1^{-1/2}\tilde{\omega}_2^{-1/2}\kappa_{2,1,\tilde{h}}\tilde{x}_2\epsilon_1$$

Since h is a symmetric function of observations 1 and 2, a U-statistic with kernel h is defined as

$$B_n'=\frac{2}{n(n-1)}\sum_{j<l}^{n}h(\hat{\omega}_{n,j,1},\tilde{X}_j,e_j,W_j;\hat{\omega}_{n,l,1},\tilde{X}_l,e_l,W_l) \qquad \text{and} \qquad B_n=\frac{1}{2}B_n'$$

A similar U-statistic is defined as

$$\tilde{B}'_n = \frac{2}{n(n-1)} \sum_{j<l}^{n} h(\omega_j, \tilde{X}_j, e_j, W_j; \omega_l, \tilde{X}_l, e_l, W_l) \qquad \text{with} \qquad \tilde{B}'_n = B'_n + o_p(1)$$

In addition, we have:

$$\begin{aligned}
E(\tilde{B}'_n) &= E(\omega_j^{-1/2}\omega_l^{-1/2}\kappa_{j,l,\tilde{h}}\tilde{X}_j e_l + \omega_j^{-1/2}\omega_l^{-1/2}\kappa_{l,j,\tilde{h}}\tilde{X}_l e_j) \\
&= 2E(\omega_j^{-1/2}\omega_l^{-1/2}\kappa_{j,l,\tilde{h}}\tilde{X}_j e_l) \\
&= 2\int_{\mathbb{R}^{q_w}} \tilde{h}^{-q_w} E[\omega_j^{-1/2}\omega_l^{-1/2}\tilde{X}_j e_l e^{it'(W_j - W_l)/\tilde{h}}] d\mu(t) \\
&= 2\int_{\mathbb{R}^{q_w}} \tilde{h}^{-q_w} E[\omega_j^{-1/2}\omega_l^{-1/2}\tilde{X}_j e^{it'W_j/\tilde{h}} e_l e^{-it'W_l/\tilde{h}}] d\mu(t) \\
&= 2\int_{\mathbb{R}^{q_w}} \tilde{h}^{-q_w} E[\omega_j^{-1/2}\tilde{X}_j e^{it'W_j/\tilde{h}}] E[\omega_l^{-1/2} e_l e^{-it'W_l/\tilde{h}}] d\mu(t) \\
&= 0 \qquad \text{since } E[\omega_l^{-1/2} e'_l e^{-it'W_l}] = 0.
\end{aligned}$$

It then follows from WLLN for U statistics that $B_n \xrightarrow{p} 0$, and we conclude that $\tilde{\beta}_{e,n}$ is a consistent estimator of β_0. Before we can derive the asymptotic distribution of $\tilde{\beta}_{e,n}$, we need to compute the variance of these U statistics. Notice first that:

$$h(\tilde{\omega}_1, \tilde{x}_1, \epsilon_1, w_1; \tilde{\omega}_2, \tilde{x}_2, \epsilon_2, w_2) = \tilde{\omega}_1^{-1/2}\tilde{\omega}_2^{-1/2}\kappa_{1,2,\tilde{h}}\tilde{x}_1\epsilon_2 + \tilde{\omega}_1^{-1/2}\tilde{\omega}_2^{-1/2}\kappa_{2,1,\tilde{h}}\tilde{x}_2\epsilon_1$$

We then have:

$$\begin{aligned}
& h_1(\tilde{\omega}_1, \tilde{x}_1, \epsilon_1, w_1) \\
\equiv\ & E(h(\omega_1, \tilde{X}_1, e_1, W_1; \omega_2, \tilde{X}_2, e_2, W_2) | \omega_1 = \tilde{\omega}_1, \tilde{X}_1 = \tilde{x}_1, e_1 = \epsilon_1, W_1 = w_1) \\
=\ & \tilde{h}^{-q_w} E[\tilde{\omega}_1^{-1/2}\omega_2^{-1/2} \int_{\mathbb{R}^{q_w}} e^{it'(w_1 - W_2)/\tilde{h}} d\mu(t) \tilde{x}_1 e_2 + \tilde{\omega}_1^{-1/2}\omega_2^{-1/2} \int_{R^{q_w}} e^{it'(W_2 - w_1)/\tilde{h}} d\mu(t) \tilde{X}_2 \epsilon_1] \\
=\ & \tilde{h}^{-q_w} E[\tilde{\omega}_1^{-1/2}\omega_2^{-1/2} \int_{\mathbb{R}^{q_w}} e^{it'(w_1 - W_2)/\tilde{h}} d\mu(t) \tilde{x}_1 e_2] + \tilde{h}^{-q_w} E[\tilde{\omega}_1^{-1/2}\omega_2^{-1/2} \int_{R^{q_w}} e^{it'(W_2 - w_1)/\tilde{h}} d\mu(t) \tilde{X}_2 \epsilon_1] \\
=\ & \tilde{h}^{-q_w} \int_{\mathbb{R}^{q_w}} \tilde{\omega}_1^{-1/2} e^{-it'w_1/\tilde{h}} \epsilon_1 E[\omega_2^{-1/2} e^{it'W_2/\tilde{h}} \tilde{X}_2] d\mu(t)
\end{aligned}$$

where the last equality follows from studying each term of the RHS separately:

$$\begin{aligned}
& \tilde{h}^{-q_w} E[\tilde{\omega}_1^{-1/2}\omega_2^{-1/2} \int_{\mathbb{R}^{q_w}} e^{it'(w_1 - W_2)/\tilde{h}} d\mu(t) \tilde{x}_1 e_2] \\
=\ & \tilde{h}^{-q_w} \int_{\mathbb{R}^{q_w}} E[\tilde{\omega}_1^{-1/2}\omega_2^{-1/2} e^{it'w_1/\tilde{h}} e^{-it'W_2/\tilde{h}} \tilde{x}_1 e_2] d\mu(t) \\
=\ & \tilde{h}^{-q_w} \tilde{\omega}_1^{-1/2} \int_{\mathbb{R}^{q_w}} e^{it'w_1/\tilde{h}} \tilde{x}_1 E[\omega_2^{-1/2} e^{-it'W_2/\tilde{h}} e_2] d\mu(t) \\
=\ & 0
\end{aligned}$$

The second term of the RHS is:

$$\begin{aligned}
& \tilde{h}^{-q_w} E[\tilde{\omega}_1^{-1/2} \omega_2^{-1/2} \int_{R^{q_w}} e^{it'(W_2-w_1)/\tilde{h}} d\mu(t) \tilde{X}_2 \epsilon_1] \\
= \; & \tilde{h}^{-q_w} \int_{\mathbb{R}^{q_w}} E[\tilde{\omega}_1^{-1/2} \omega_2^{-1/2} e^{it'W_2/\tilde{h}} e^{-it'w_1/\tilde{h}} \tilde{X}_2 \epsilon_1] d\mu(t) \\
= \; & \tilde{h}^{-q_w} \int_{\mathbb{R}^{q_w}} \tilde{\omega}_1^{-1/2} e^{-it'w_1/\tilde{h}} \epsilon_1 E[\omega_2^{-1/2} e^{it'W_2/\tilde{h}} \tilde{X}_2] d\mu(t)
\end{aligned}$$

Since $E(h_1(\omega_1, \tilde{X}_1, e_1, W_1)) = 0$, we have:

$$\begin{aligned}
Var[h_1(\omega_1, \tilde{X}_1, e_1, W_1)] &= Var[\tilde{h}^{-q_w} \int_{\mathbb{R}^{q_w}} \tilde{\omega}_1^{-1/2} e^{-it'W_1/\tilde{h}} \epsilon_1 E[\omega_2^{-1/2} e^{it'W_2/\tilde{h}} \tilde{X}_2] d\mu(t)] \\
&= Var[\tilde{\omega}_1^{-1/2} \epsilon_1 E[\omega_2^{-1/2} \kappa_{2,1,\tilde{h}} \tilde{X}_2 | 1]] \\
&= E[\omega_1^{-1} e_1^2 E[\omega_2^{-1/2} \kappa_{2,1,\tilde{h}} \tilde{X}_2 | 1] E[\omega_2^{-1/2} \kappa_{2,1,\tilde{h}} \tilde{X}_2' | 1]] \\
&= E[\omega_1^{-1} e_1^2 \omega_1^{-1/2} E(\tilde{X}_1 | W_1) f(W_1) \omega_1^{-1/2} E(\tilde{X}_1' | W_1) f(W_1)] \\
&= E[\omega_1^{-1} e_1^2 \omega_1^{-1} E(\tilde{X}_1 | W_1) E(\tilde{X}_1' | W_1) f(W_1)^2] \\
&= E[Var[e_1 | W_1]^{-2} f(W_1)^{-2} e_1^2 E(\tilde{X}_1 | W_1) E(\tilde{X}_1' | W_1) f(W_1)^2] \\
&= E[Var[e_1 | W_1]^{-2} e_1^2 E(\tilde{X}_1 | W_1) E(\tilde{X}_1' | W_1)] \\
&= E[Var[e_1 | W_1]^{-2} E(e_1^2 | W_1) E(\tilde{X}_1 | W_1) E(\tilde{X}_1' | W_1)] \\
&= E[Var[e_1 | W_1]^{-1} E(\tilde{X}_1 | W_1) E(\tilde{X}_1' | W_1)] \\
&= E[Var[e_j(\beta_0) | W_j]^{-1} E(\tilde{X}_j | W_j) E(\tilde{X}_j' | W_j)]
\end{aligned}$$

Following Serfling (1980) (section 5.5.1), it follows that:

$$\sqrt{n}(B_n' - 0) \xrightarrow{d} \mathcal{N}(0, 4Var[h_1(\omega_1, \tilde{X}_1, e_1, W_1)])$$

And, as a result, we have:

$$\sqrt{n}(\tilde{\beta}_{e,n} - \beta_0) \xrightarrow{d} \mathcal{N}(0, [E[Var[e_j(\beta_0) | W_j]^{-1} E(\tilde{X}_j | W_j) E(\tilde{X}_j' | W_j)]]^{-1}) .$$

□

Proof of Theorem 2.5

• Preliminary result #3:
Under the assumptions of Theorem 2.5,

$$C_n \equiv \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{l \neq j}^{n} \hat{\omega}_{n,j,1}^{-1/2} \hat{\omega}_{n,l,1}^{-1/2} \kappa_{j,l,\tilde{h}} \hat{\tilde{X}}_j \hat{\tilde{X}}_l' \xrightarrow{P} A$$

Proof. Recall that

$$\begin{aligned}
\widehat{\tilde{X}}_j &= X_j - \hat{g}_x(Z_j) \\
&= X_j - E(X_j|Z_j) + E(X_j|Z_j) - \hat{g}_x(Z_j) \\
&= X_j - E(X_j|Z_j) + \mathcal{O}_p(v_n) \\
&= \tilde{X}_j + \mathcal{O}_p(v_n)
\end{aligned}$$

We then have:

$$\begin{aligned}
A_n - C_n &= \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{l \neq j}^{n} \hat{\omega}_{n,j,1}^{-1/2} \hat{\omega}_{n,l,1}^{-1/2} \kappa_{j,l,\tilde{h}} \tilde{X}_j \tilde{X}_l' \\
&\quad - \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{l \neq j}^{n} \hat{\omega}_{n,j,1}^{-1/2} \hat{\omega}_{n,l,1}^{-1/2} \kappa_{j,l,\tilde{h}} [\tilde{X}_j + O_p(v_n)][\tilde{X}_l + O_p(v_n)]' \\
&= \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{l \neq j}^{n} \hat{\omega}_{n,j,1}^{-1/2} \hat{\omega}_{n,l,1}^{-1/2} \kappa_{j,l,\tilde{h}} \tilde{X}_j \tilde{X}_l' \\
&\quad - \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{l \neq j}^{n} \hat{\omega}_{n,j,1}^{-1/2} \hat{\omega}_{n,l,1}^{-1/2} \kappa_{j,l,\tilde{h}} [\tilde{X}_j \tilde{X}_l' + O_p(v_n) \tilde{X}_l' + \tilde{X}_j O_p(v_n) + O_p(v_n^2)] \\
&= \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{l \neq j}^{n} \hat{\omega}_{n,j,1}^{-1/2} \hat{\omega}_{n,l,1}^{-1/2} \kappa_{j,l,\tilde{h}} [O_p(v_n) \tilde{X}_l' + \tilde{X}_j O_p(v_n) + O_p(v_n^2)] \\
&= \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{l \neq j}^{n} [\omega_j + \mathcal{O}_p\left(v_{b,n}\right) + o_p(1)]^{-1/2} \hat{\omega}_{n,l,1}^{-1/2} \kappa_{j,l,\tilde{h}} [O_p(v_n) \tilde{X}_l' + \tilde{X}_j O_p(v_n) + O_p(v_n^2)] \\
&\xrightarrow{P} 0
\end{aligned}$$

And the result follows from Preliminary result #2 in the previous subsection. □

- We now show the consistency of $\hat{\beta}_{e,n}$.

Proof. Recall that

$$\begin{aligned}
\widehat{\tilde{y}}_i &= y_i - \hat{g}_y(Z_i) \\
&= y_i - E(y_i|Z_i) + E(y_i|Z_i) - \hat{g}_y(Z_i) \\
&= (X_i - \hat{g}_X(Z_i))'\beta_0 + (\hat{g}_X(Z_i) - E(X_i|Z_i))'\beta_0 + e_i + E(y_i|Z_i) - \hat{g}_y(Z_i) \\
&= \widehat{\tilde{X}}_i'\beta_0 + (\hat{g}_X(Z_i) - E(X_i|Z_i))'\beta_0 + e_i + E(y_i|Z_i) - \hat{g}_y(Z_i)
\end{aligned}$$

The estimator $\hat{\beta}_{e,n}$ can then be rewritten as:

$$\begin{aligned}
\hat{\beta}_{e,n} &= [C_n]^{-1}[\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\hat{\tilde{X}}_j[\hat{\tilde{X}}_l'\beta_0 \\
&\quad +(\hat{g}_X(Z_l)-E(X_l|Z_l))'\beta_0+e_l+E(y_l|Z_l)-\hat{g}_y(Z_l)]] \\
&= \beta_0+[C_n]^{-1}[\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\hat{\tilde{X}}_j[(\hat{g}_X(Z_l)-E(X_l|Z_l))'\beta_0 \\
&\quad +e_l+E(y_l|Z_l)-\hat{g}_y(Z_l)]] \\
&= \beta_0+[C_n]^{-1}D_n \qquad (2.5)
\end{aligned}$$

with

$$D_n \equiv \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\hat{\tilde{X}}_j[(\hat{g}_X(Z_l)-E(X_l|Z_l))'\beta_0+e_l+E(y_l|Z_l)-\hat{g}_y(Z_l)]$$

which is decomposed into $D_n = E_n + F_n + G_n$ after introducing

$$\begin{aligned}
E_n &\equiv \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\hat{\tilde{X}}_j(\hat{g}_X(Z_l)-E(X_l|Z_l))'\beta_0 \\
F_n &\equiv \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\hat{\tilde{X}}_j e_l \\
G_n &\equiv \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\hat{\tilde{X}}_j[E(y_l|Z_l)-\hat{g}_y(Z_l)]
\end{aligned}$$

To obtain the consistency of $\hat{\beta}_{e,n}$, we show that the probability limits for E_n, F_n, and G_n are all zero.

$$\begin{aligned}
E_n &= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\hat{\tilde{X}}_j(\hat{g}_X(Z_l)-E(X_l|Z_l))'\beta_0 \\
&= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}[\tilde{X}_j+O_p(v_n)](O_p(v_n))'\beta_0 \\
&\xrightarrow{P} 0
\end{aligned}$$

$$\begin{aligned}
F_n &= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}[\tilde{X}_j+O_p(v_n)]e_l \\
&= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}[\tilde{X}_j e_l+O_p(v_n)e_l] \\
&\xrightarrow{P} 0
\end{aligned}$$

since, from the proof of Proposition 5, $E(\omega_j^{-1/2}\omega_l^{-1/2}\kappa_{j,l,\tilde{h}}\tilde{X}_j e_l) = 0$.

$$\begin{aligned} G_n &= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\hat{\tilde{X}}_j[E(y_l|Z_l)-\hat{g}_y(Z_l)] \\ &= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}[\tilde{X}_j+O_p(v_n)][O_p(v_n)] \\ &\overset{P}{\to} 0 \end{aligned}$$

And the consistency of $\hat{\beta}_{e,n}$ follows. □

- We now derive the asymptotic distribution of $\hat{\beta}_{e,n}$.

Proof. From equation (2.5) and preliminary result #3, we have:

$$\sqrt{n}(\hat{\beta}_{e,n}-\beta_0) = [C_n]^{-1}\sqrt{n}[E_n+F_n+G_n] \qquad \text{with} \qquad \texttt{Plim}C_n = A\,.$$

We now study each term separately.

$$\begin{aligned} \sqrt{n}F_n &= \sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}[\tilde{X}_j+O_p(v_n)]e_l \\ &= \sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}[\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\tilde{X}_j e_l+\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}O_p(v_n)e_l] \\ &= \frac{\sqrt{n}}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\tilde{X}_j e_l+\frac{\sqrt{n}}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}O_p(v_n)e_l \\ &= \sqrt{n}B_n+\sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}O_p(v_n)e_l \\ &\quad \text{with } B_n \text{ defined in equation (2.4) above} \\ &= \sqrt{n}B_n+\mathcal{O}_p(v_n) \end{aligned}$$

which follows from a CLT applied to the 2nd term on the RHS which is a U-statistic with mean 0.

$$\sqrt{n}E_n = \sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\tilde{X}_j(\hat{g}_X(Z_l)-E(X_l|Z_l))'\beta_0+o_p(1)$$

Treating Z_l as a nonrandom vector, and following Li and Racine (page 63), we have:

$$\sqrt{nh_1...h_{q_z}}(\hat{g}_X(Z_l)-E(X_l|Z_l)-\sum_{s=1}^{q_z}h_s^2 B_s(Z_l)) \overset{d}{\to} N(0,\zeta^{q_z}\sigma^2(Z_l)/f(Z_l)) \tag{2.6}$$

where $\zeta = \int k(v)^2dv$, $B_s(Z_l) = \frac{\int v^2k(v)dv}{2}\{2f_s(Z_l)E_s(X_l|Z_l)+f(Z_l)E_{ss}(X_l|Z_l)\}/f(Z_l)$, $\sigma^2(Z_l) = E[u_{X,i}^2|Z_l]$ and $r_s(Z_l)$ and $r_{ss}(Z_l)$ are the first and second order derivatives of $r(Z_l)$ with respect to $Z_{s,l}$. It then follows that:

$$\sqrt{n}E_n \quad = \quad \sqrt{n}O_p(\frac{1}{\sqrt{nh_1...h_{q_z}}}) + \sqrt{n}R(n,j,l) + o_p(1)$$

where $R(n,j,l) = \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\tilde{X}_j(\sum_{s=1}^{q_z}h_s^2 B_s(Z_l))'\beta_0$.

$$\sqrt{n}G_n \quad = \quad \sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\hat{\tilde{X}}_j[E(y_l|Z_l) - \hat{g}_y(Z_l)]$$

Similarly to equation (2.6) above, we have:

$$\sqrt{nh_1...h_{q_z}}(E(y_l|Z_l) - \hat{g}_y(Z_l) + \sum_{s=1}^{q_z}h_s^2 B_s'(Z_l)) \xrightarrow{d} N(0, \zeta^{q_z}\sigma'^2(Z_l)/f(Z_l))$$

where $B_s'(Z_l) = \frac{\int v^2k(v)dv}{2}\{2f_s(Z_l)E_s(y_l|Z_l) + f(Z_l)E_{ss}(y_l|Z_l)\}/f(Z_l)$ and $\sigma'^2(Z_l) = E[u_{Y,i}^2|Z_l]$. And it follows that:

$$\sqrt{n}G_n \quad = \quad \sqrt{n}O_p(\frac{1}{\sqrt{nh_1...h_{q_z}}}) - \sqrt{n}R'(n,j,l) + o_p(1)$$

where $R'(n,j,l) = \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\hat{\omega}_{n,j,1}^{-1/2}\hat{\omega}_{n,l,1}^{-1/2}\kappa_{j,l,\tilde{h}}\tilde{X}_j(\sum_{s=1}^{q_z}h_s^2 B_s'(Z_l))$.

Ultimately, we have:

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_{e,n} - \beta_0) \quad &= \quad [C_n]^{-1}\left[\sqrt{n}F_n + \sqrt{n}E_n + \sqrt{n}G_n\right] \\
&= \quad [C_n]^{-1}\Bigg[\sqrt{n}B_n + \sqrt{n}(O_p(\frac{1}{\sqrt{nh_1...h_{q_z}}}) + R(n,j,l)) \\
&\qquad + \sqrt{n}\left(O_p(\frac{1}{\sqrt{nh_1...h_{q_z}}}) - R'(n,j,l)\right) + o_p(1)\Bigg] \\
&= \quad [C_n]^{-1}[\sqrt{n}B_n + \sqrt{n}(o_p(1) + R(n,j,l) - R'(n,j,l)) + o_p(1)]
\end{aligned}$$

which implies that

$$\sqrt{n}\left(\hat{\beta}_{e,n} - \beta_0 - C_n^{-1}E[R(n,j,l) - R'(n,j,l)]\right) = C_n^{-1}[\sqrt{n}B_n] + o_p(1) \qquad (2.7)$$

and the result follows. □

3 Simulation study: Results for the L-L model

In this section, we consider the L-L model with a sample of size $n = 200$ and $5,000$ Monte-Carlo replications generated as follows,

$$\begin{aligned} X_i^* &= 2W_i + Z_i + v_i \\ X_i &= 8\text{scale}(X_i^*) \\ y_i &= 2X_i + 3Z_i + e_i \end{aligned}$$

In this fully linear model, we expect all the estimators we consider to behave and perform quite similarly. This is confirmed with the results displayed in Table A.1. The performance of these four estimators is quite similar with respect to bias, standard deviations and rejection rates.

In Table A.2, we display the performance of R-SMD for different bandwidths. The results do not change much compared to those in Table A.1, where the bandwidth was chosen using the rule of thumb. In Table A.3, we display the performance of R-SMD for different choices of $\mu(.)$ (and $k(.)$). Once again, the results are very similar for the different choices of $\mu(.)$.

Estimator	R-SMD	R-GMM		Sieves-GMM		GMM	
		W	(W, W^2)	W	(W, W^2)	W	(W, W^2)
Bias	0.005	0.006	0.007	0.001	0.001	0.000	0.000
SE	0.018	0.016	0.016	0.015	0.015	0.015	0.015
Asympt. Homosk. SE	0.021	0.019	0.019	0.015	0.015	0.015	0.015
Asympt. Heterosk. SE	0.020	0.018	0.018	0.015	0.015	0.015	0.015
Rej. rate for Homosk. SE	0.034	0.033	0.035	0.059	0.061	0.054	0.056
Rej. rate for Heteros. SE	0.036	0.040	0.043	0.063	0.063	0.061	0.061

Table A.1: Simulation Results for the L-L model using $n = 200$ and $M = 5,000$

We report the Monte-Carlo bias and Monte-Carlo standard error (SE), as well as the average of the asymptotic SE assuming either homoskedasticity or heteroskedasticity, and the empirical rejection rates of the null hypothesis $H_0 : \beta = \beta_0$ using a 5% t-test for the R-SMD, R-GMM, Sieves-GMM and GMM estimators. GMM, Sieves-GMM, and R-GMM are computed using either one moment with instrument W or two moments with instruments W and W^2.

Bandwidth	Bias	SE	Asymp. Heterosk. SE	Rej. rate
0.08705506	-0.002	0.019	0.022	0.025
0.1519344	-0.001	0.018	0.021	0.025
0.2168137	0.000	0.018	0.021	0.025
0.2816931	0.002	0.018	0.021	0.028
0.3465724	0.005	0.018	0.020	0.036

Table A.2: Simulation Results for the L-L model using $n = 200$ and $M = 5,000$

We report the Monte-Carlo bias and Monte-Carlo standard error (SE), as well as the average of the asymptotic heteroskedasticity-robust SE, and the empirical rejection rates of the null hypothesis $H_0 : \beta = \beta_0$ using a 5% t-test for the R-SMD as a function of the bandwidth used in the Nadaraya-Watson estimate.

R-SMD Estimator	Gaussian $\mu(.)$ (Gaussian $k(.)$)	Cauchy $\mu(.)$ (Laplace $k(.)$)	$\texttt{sinc}^2$ $\mu(.)$ (Triangular $k(.)$)
Bias	0.005	0.005	0.005
SE	0.018	0.018	0.020
Asympt. Homosk. SE	0.021	0.021	0.023
Asympt. Heterosk. SE	0.020	0.021	0.022
Rej. rate for Homosk. SE	0.034	0.034	0.039
Rej. rate for Heterosk. SE	0.036	0.036	0.040

Table A.3: Simulation results for the L-L model with $n = 200$ and $M = 5,000$

We report the Monte-Carlo bias and Monte-Carlo standard error (SE), the average of the asymptotic heteroskedasticity-robust SE, and the empirical rejection rates of the null hypothesis $H_0 : \beta = \beta_0$ using a 5% t-test for three R-SMD estimators associated with $\mu(.)$ chosen as the CDF of: (i) a standard Gaussian distribution, (ii) a Cauchy distribution; (iii) a *sinc*^{2} distribution.

4 Empirical application

• **Alternative set of control variables:**

The following tables correspond to Tables 3 and 4 in the main paper using the full set of control variables (see Table 3 in Dinkelman (2011)).

Panel A: Effects on employment

Outcome is Δ_t in	IV (L-L)	IV (Q-L)	R-SMD (NL-NL)
Female employment rate	0.095*	0.057	0.061
	(0.051)	(0.043)	(0.044)
Male employment rate	0.035	-0.012	0.074
	(0.062)	(0.057)	(0.061)

Panel B: Effects on Household energy sources & other household services

Outcome is Δ_t in	IV (L-L)	IV (Q-L)	R-SMD (NL-NL)
Lighting with electricity	0.635***	0.358***	0.386***
	(0.176)	(0.113)	(0.146)
Cooking with wood	-0.275**	-0.216**	-0.191
	(0.123)	(0.095)	(0.122)
Cooking with electricity	0.228***	0.128**	0.155**
	(0.077)	(0.051)	(0.074)
Water nearby	-0.372*	-0.363**	-0.626**
	(0.197)	(0.167)	(0.246)
Flush toilet	0.067	0.069	0.104*
	(0.055)	(0.052)	(0.060)

Table A.4: Impact of electrification on Employment (Panel A) and on Household energy sources & other household services (Panel B)

Note: *** Significant at 1%, ** at 5%, * at 10%. Each cell in the table presents estimates of the Eskom project variable coefficient (and robust standard error) from an IV regression of the dependent variable on the Eskom project indicator and control variables (that include baseline controls and district fixed effects, and two additional variables for access to water and toilet, except for the last 2 rows of Panel B where these 2 controls are omitted; see Table 3 in Dinkelman (2011)). In Panel A, the dependent variable is the change in female (or male) employment rate between 1996 and 2001; in Panel B, the outcome variables measure the change in fraction of households using different energy sources or with access to basic services. Each regression contains $N = 1,816$ except for change in fraction of households using wood which contains $N = 1,807$ due to missing data on this variable.

Panel A								
Δ_t log population			Δ_t Females with High School			Δ_t Males with High School		
IV (L-L)	IV (Q-L)	R-SMD (NL-NL)	IV (L-L)	IV (Q-L)	R-SMD (NL-NL)	IV (L-L)	IV (Q-L)	R-SMD (NL-NL)
3.897***	3.174***	2.780***	0.130**	0.103**	0.171**	0.077	0.048	0.071
(1.416)	(0.975)	(0.800)	(0.058)	(0.049)	(0.070)	(0.050)	(0.042)	(0.050)
Panel B								
Δ_t log non in-migrant population			Δ_t Females empl. excl. in-migrants			Δ_t Males empl. excl. in-migrants		
IV (L-L)	IV (Q-L)	R-SMD (NL-NL)	IV (L-L)	IV (Q-L)	R-SMD (NL-NL)	IV (L-L)	IV (Q-L)	R-SMD (NL-NL)
4.349***	3.416***	3.043***	0.116*	0.076	0.056	0.086	0.025	0.096
(1.573)	(1.038)	(0.866)	(0.068)	(0.048)	(0.042)	(0.069)	(0.055)	(0.061)

Table A.5: Impact of electrification on Population growth, skill composition of labor force and employment of incumbents

Note: *** Significant at 1%, ** at 5%, * at 10%. Each cell in the table presents estimates of the Eskom project variable coefficient (and robust standard error) from an IV regression of the dependent variable on the Eskom project indicator and control variables (that include baseline controls and district fixed effects; see Table 3 in Dinkelman (2011)). Dependent variable in panel A, column 1, is change in log African population; in columns 2-3 it is the change in fraction of women or men that have a completed high school education. Dependent variable in panel B, column 1, is the change in log African non–in-migrant population where in-migrants have been subtracted from the total number of adults in the community in each year. In columns 2-3 of panel B, the outcomes are change in female and male employment rates where the employment variables exclude the number of in-migrants to each community in each year. Each regression contains $N = 1,816$. Note that the results for IV (L-L) and IV (Q-L) are taken from Dieterle and Snell (2016).

- **Alternative estimation strategy:**

This Table corresponds to Panel A, row 2 of Table 1.3 and Panel A, columns 2 and 3 of Table 1.4 in the main text where we report estimates of the effect of electrification on male employment rate and on the change in fraction of women (men) that have a completed high school education using R-GMM (with up to 3 powers of W as instruments); for comparison purposes, we also report estimates obtained using R-SMD.

Panel A: male employment rate			
R-GMM (W)	R-GMM (W, W^2)	R-GMM (W, W^2, W^3)	R-SMD
0.107*	0.057	0.048	0.061
(0.062)	(0.045)	(0.041)	(0.059)
Panel B: fraction of women that have a completed high school education			
R-GMM (W)	R-GMM (W, W^2)	R-GMM (W, W^2, W^3)	R-SMD
0.196***	0.145***	0.144***	0.195**
(0.063)	(0.046)	(0.047)	(0.082)
Panel C: fraction of men that have a completed high school education			
R-GMM (W)	R-GMM (W, W^2)	R-GMM (W, W^2, W^3)	R-SMD
0.118**	0.076*	0.070	0.089
(0.049)	(0.040)	(0.039)	(0.057)

Table A.6: Impact of electrification on male employment rate (Panel A) and on the change in fraction of women (and men) that have a completed high school education (Panels B and C)

Note: *** Significant at 1%, ** at 5%, * at 10%. Each cell in the table presents estimates of the Eskom project variable coefficient (and robust standard error) from an IV regression of the dependent variable on the Eskom project indicator and control variables (that include baseline controls and district fixed effects; see Table 3 in Dinkelman (2011)). In Panel A, the dependent variable is the change in male employment rate between 1996 and 2001; in Panels B and C, the outcome variables measure the change in fraction of women and men that have a completed high school education. Each regression contains $N = 1,816$.

- **Identification-robust inference:**

This Table corresponds to Panel A of Table 1.3 in the main text where we report identification-robust 95% confidence intervals on the effect of electrification on male and female employment rate obtained with Anderson-Rubin and Conditional Likelihood Ratio, as well as the F-test statistic which is often used as a rule-of-thumb to evaluate weak identification. We consider both the set of controls (as in Table 3 Panel A), as well as the restricted set obtained by selecting the first two PCA as explained on p12. For comparison purposes, we also report the corresponding (non-robust) confidence intervals. We use the R package `ivmodel`.

Panel A: Effects on employment (full set of controls)

Outcome is Δ_t in	IV (L-L)	IV (Q-L)
Female employment rate	0.090*	0.053
IV	[-0.008,0.188]	[-0.031,0.137]
AR	[-0.004,0.255]	[-0.025,0.185]
CLR	[-0.004,0.255]	[-0.032,0.199]
Male employment rate	0.033	-0.013
IV	[-0.089,0.155]	[-0.125,0.099]
AR	[-0.100,0.199]	[-0.145,0.115]
CLR	[-0.100,0.199]	[-0.157,0.126]
F-test statistic	13.7	8.9

Panel B: Effects on employment (restricted set of controls)

Outcome is Δ_t in	IV (L-L)	IV (Q-L)
Female employment rate	0.115**	0.039
IV	[0.013,0.218]	[-0.023,0.100]
AR	[0.024,0.301]	[0.012,0.087]
CLR	[0.024,0.301]	[-0.023,0.137]
Male employment rate	0.128	0.040
IV	[-0.007,0.263]	[-0.049,0.128]
AR	[0.006,0.362]	[-0.024,0.129]
CLR	[0.006,0.362]	[-0.046,0.161]
F-test statistic	13.3	14.1

Table A.7: Impact of electrification on Employment measured as the change in female (or male) employment rate between 1996 and 2001

Note: *** Significant at 1%, ** at 5%, * at 10%. Each cell in the table presents estimates of the Eskom project variable coefficient (and robust standard error) from an IV regression of the dependent variable on the Eskom project indicator and control variables: in Panel A, we include baseline controls and district fixed effects; in Panel B, we include the first two PCA on the set of controls included in Panel A. Each regression contains $N = 1,816$.

Appendix B

Estimation of Heterogeneous Treatment Effects Using a Conditional Moment Based Approach

1 Tables for the Returns of Education

Homogeneous treatment effects				
Estimator	Z_1	Z_2	(Z_1, Z_1X_1)	(Z_1, Z_2)
GMM	0.138**	0.223**	0.138**	0.144***
	(0.055)	(0.092)	(0.055)	(0.055)
GMM-Lasso	0.142	0.244	0.156	0.064
	(0.108)	(0.207)	(0.111)	(0.082)
DRSMD-Lasso	0.187*	0.574	0.128*	0.203**
	(0.104)	(0.728)	(0.078)	(0.098)
DRSMD-2SOLS	0.214	0.738	0.137*	0.234**
	(0.131)	(1.226)	(0.075)	(0.121)

Table B.1: Homogeneous Treatment Effects of years of education on the log wage

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used. Every regression contains 3010 observations.

Homogeneous treatment effects				
Estimator	Z_1	Z_2	(Z_1, Z_1X_1)	(Z_1, Z_2)
GMM	0.788**	1.025**	0.656**	0.867**
	(0.349)	(0.447)	(0.309)	(0.349)
GMM-Lasso	0.793	1.085	1.017	0.648
	(0.676)	(0.967)	(0.715)	(0.609)
DRSMD-Lasso	1.757	2.474	0.914	1.394*
	(1.457)	(2.646)	(0.620)	(0.820)
DRSMD-2SOLS	2.194	3.677	0.985	1.642
	(2.269)	(5.765)	(0.728)	(1.084)

Table B.2: Homogeneous Treatment Effects of education on the log wage

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used. Every regression contains 3010 observations.

Heterogeneous treatment effects				
Estimator for θ_{w0}	Z_1	Z_2	(Z_1, Z_1X_1)	(Z_1, Z_2)
GMM			0.137	-0.524
			(0.085)	(0.659)
GMM-Lasso			-0.095	-0.547
			(0.202)	(0.732)
DRSMD-Lasso	0.187	0.095	0.382**	0.363**
	(0.124)	(0.079)	(0.167)	(0.181)
DRSMD-2SOLS	0.149*	-0.022	0.192**	0.190**
	(0.083)	(0.092)	(0.090)	(0.093)
Estimator for θ_{wx0}	Z_1	Z_2	(Z_1, Z_1X_1)	(Z_1, Z_2)
GMM			0.000	0.064
			(0.006)	(0.063)
GMM-Lasso			0.024	0.070
			(0.018)	(0.083)
DRSMD-Lasso	0.000	0.030	-0.026	-0.015
	(0.002)	(0.035)	(0.017)	(0.011)
DRSMD-2SOLS	0.006	0.043	-0.007	0.004
	(0.004)	(0.047)	(0.011)	(0.008)

Table B.3: Heterogeneous Treatment Effects of years of education on the log wage

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used. Every regression contains 3010 observations.

Heterogeneous treatment effects				
Estimator for $\theta_{w0} + \theta_{wx0}E(X)$	Z_1	Z_2	(Z_1, Z_1X_1)	(Z_1, Z_2)
GMM			0.138**	0.129
			(0.055)	(0.083)
GMM-Lasso			0.149	0.164
			(0.116)	(0.172)
DRSMD-Lasso	0.187*	0.403	0.119	0.216**
	(0.104)	(0.349)	(0.080)	(0.107)
DRSMD-2SOLS	0.209*	0.414	0.124	0.228**
	(0.124)	(0.388)	(0.084)	(0.112)

Table B.4: Heterogeneous Treatment Effects of years of education on the log wage

Note: Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used. Every regression contains 3010 observations.

2 Proofs of the Theoretical Results

2.1 Equivalence between the Objective Functions (2.5) and (2.6)

Proof. The objective function (2.5) can be written as

$$M_\infty(\theta, g) = \int_{\mathbb{R}^{q_z}} E[\epsilon_j(\theta, g)e^{it'Z_j}]E[\epsilon_l(\theta, g)e^{-it'Z_l}]d\mu(t)$$

From Assumption 4(vi), for all $j \neq l$, $Cov(\epsilon_j(\theta, g)e^{it'Z_j}, \epsilon_l(\theta, g)e^{-it'Z_l}) = 0$. Thus, for all $j \neq l$, we have:

$$\begin{aligned} M_\infty(\beta) &= \int_{\mathbb{R}^{q_z}} E(\epsilon_j(\theta, g)e^{it'Z_j}\epsilon_l(\theta, g)e^{-it'Z_l})d\mu(t) \\ &= \int_{\mathbb{R}^{q_z}} E(\epsilon_j(\theta, g)\epsilon_l(\theta, g)e^{it'(Z_j - Z_l)})d\mu(t) \\ &= E(\int_{\mathbb{R}^{q_z}} \epsilon_j(\theta, g)\epsilon_l(\theta, g)e^{it'(Z_j - Z_l)}d\mu(t)) \end{aligned}$$

Thus, the objective function becomes

$$M_\infty(\beta) = E(\epsilon_j(\theta, g)\epsilon_l(\theta, g)\kappa_{j,l})$$

where $\kappa_{j,l} = k(Z_j - Z_l) = \int_{\mathbb{R}^{q_z}} e^{it'(Z_j - Z_l)}d\mu(t)$. And $k(u)$ is the inverse Fourier transform of $d\mu(t)$ with $u = Z_j - Z_l$. □

2.2 Proof of Proposition 3

The orthogonal FOC in Equation (2.9) implies that

$$E\left[\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)(\tilde{y}_l - \tilde{P}_l'\theta)\kappa_{j,l}\right] = 0$$

$$E\left[\kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)\tilde{y}_l - \kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)\tilde{P}_l'\theta\right] = 0$$

$$\theta_0 = E\left[\kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)\tilde{P}_l'\right]^{-1} E\left[\kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)\tilde{y}_l\right]$$

The proof for invertibility of $E\left[\kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)\tilde{P}_l'\right]$ follows the same steps of proofs for identification in Antoine and Sun (2021).

2.3 Proof of Orthogonal Properties of Equation (2.7) and (2.9)

In this section, we check the orthogonal properties of the two FOCs.

$$M_\infty(\theta, g) = -\frac{1}{2}E[\epsilon_j(\theta)\epsilon_l(\theta)\kappa_{j,l}] = -\frac{1}{2}E[(\tilde{y}_j - \tilde{P}_j'\theta)(\tilde{y}_l - \tilde{P}_l'\theta)\kappa_{j,l}]$$

where g stands for all nuisance parameters.

The FOC defined in Equation (2.7) is not orthogonal

The key parameter is θ, the true value is θ_0. The nuisance parameters are $g_P(X_i)$ and $g_y(X_i)$. Their true values are $E(P_i|X_i) = g_{0,P}(X_i)$ and $E(y_i|X_i) = g_{0,y}(X_i)$.

The first order condition is written as follows:

$$E[(P_j - g_P(X_j))(y_l - g_y(X_l) - [P_l - g_P(X_l)]'\theta)\kappa_{j,l}] = 0$$

If $P_i = [W_i, f(W_i, X_i)]'$ and there are only one variable in X_i, $\partial_\theta\varphi(D;\theta,g)$ becomes:

$$\partial_\theta\varphi(D;\theta,g) = \begin{bmatrix} (P_{1j} - g_{P_1}(X_j))[y_l - g_y(X_l) - (P_{1l} - g_{P_1}(X_l))\theta_1 - (P_{2l} - g_{P_2}(X_l))\theta_2]\kappa_{j,l} \\ (P_{2j} - g_{P_2}(X_j))[y_l - g_y(X_l) - (P_{1l} - g_{P_1}(X_l))\theta_1 - (P_{2l} - g_{P_2}(X_l))\theta_2]\kappa_{j,l} \end{bmatrix}$$

Prove that $\partial_g E[\partial_\theta\varphi(D;\theta_0,g_0)] \neq 0$.

Proof. The first row of $\partial_\theta\varphi(D;\theta_0, g_0 + r(g-g_0))$ is defined as I where

$$\begin{aligned} I = \quad & [P_{1j} - g_{0,P_1}(X_j) - r(g_{P_1}(X_j) - g_{0,P_1}(X_j))] \\ & [y_l - g_{0,y}(X_l) - r(g_y(X_l) - g_{0,y}(X_l)) \\ - \quad & (P_{1l} - g_{0,P_1}(X_l) - r(g_{P_1}(X_l) - g_{0,P_1}(X_l)))\theta_1 \\ - \quad & (P_{2l} - g_{0,P_2}(X_l) - r(g_{P_2}(X_l) - g_{0,P_2}(X_l)))\theta_2]\kappa_{j,l} \end{aligned}$$

According to the definition for $\partial_g E[\partial_\theta\varphi(D;\theta_0,g_0)] \neq 0$ in Chernozhukov et al. (2018), to show that $\partial_g E[\partial_\theta\varphi(D;\theta_0,g_0)] \neq 0$ we need to show $\partial_r E[I]|_{r=0} \neq 0$.

$$\partial_r E[I]|_{r=0} = -I_1 - I_2 + I_3 + I_4$$

$$\begin{aligned} I_1 &= E\left[(g_{P_1}(X_j) - g_{0,P_1}(X_j))(y_l - g_{0,y}(X_l) - (P_{1l} - g_{0,P_1}(X_l))\theta_{0,1} - (P_{2l} - g_{0,P_2}(X_l))\theta_{0,2})\kappa_{j,l}\right] \\ &= E\left[(g_{P_1}(X_j) - g_{0,P_1}(X_j))\epsilon_l\kappa_{j,l}\right] = 0 \\ I_2 &= E\left[(P_{1j} - g_{0,P_1}(X_j))(g_y(X_l) - g_{0,y}(X_l))\kappa_{j,l}\right] \neq 0 \\ I_3 &= E\left[(P_{1j} - g_{0,P_1}(X_j))(g_{P_1}(X_l) - g_{0,P_1}(X_l))\theta_1\kappa_{j,l}\right] \neq 0 \\ I_4 &= E\left[(P_{1j} - g_{0,P_1}(X_j))(g_{P_2}(X_l) - g_{0,P_2}(X_l))\theta_2\kappa_{j,l}\right] \neq 0 \end{aligned}$$

Hence, $\partial_g E[\partial_\theta\varphi(D;\theta_0,g_0)] \neq 0$

□

The FOC defined in Equation (2.9) is orthogonal

With $E[(P_{1m} - g_{0,P_1}(X_m))\kappa_{m,l}|X_l] = g_{0,\tilde{P}_{1,j}}(X_l)$, $E[\kappa_{m,l}|X_l] = g_{0,\kappa_{m,l}}(X_l)$,

$\Psi(D;\theta,g)$

$$= \begin{bmatrix} \left(P_{1j} - g_{P_1}(X_j) - \frac{g_{\tilde{P}_{1,j}}(X_l)}{g_{\kappa_{j,l}}(X_l)}\right)[y_l - g_y(X_l) - (P_{1l} - g_{P_1}(X_l))\theta_1 - (P_{2l} - g_{P_2}(X_l))\theta_2]\kappa_{j,l} \\ \left(P_{2j} - g_{P_2}(X_j) - \frac{g_{\tilde{P}_{2,j}}(X_l)}{g_{\kappa_{j,l}}(X_l)}\right)[y_l - g_y(X_l) - (P_{1l} - g_{P_1}(X_l))\theta_1 - (P_{2l} - g_{P_2}(X_l))\theta_2]\kappa_{j,l} \end{bmatrix}$$

Prove that $\partial_g E[\Psi(D;\theta_0,g_0)] = 0$.

Proof. The first row of $\Psi(D;\theta_0,g_0+r(g-g_0))$ is I'.

$$\begin{aligned} I' = & \quad \left(P_{1j} - g_{0,P_1}(X_j) - r(g_{P_1}(X_j) - g_{0,P_1}(X_j)) - \tfrac{g_{0,\tilde{P}_{1,j}}(X_l)+r[g_{\tilde{P}_{1,j}}(X_l)-g_{0,\tilde{P}_{1,j}}(X_l)]}{g_{0,\kappa_{m,l}}(X_l)+r[g_{\kappa_{j,l}}(X_l)-g_{0,\kappa_{m,l}}(X_l)]}\right) \\ & \qquad [y_l - g_{0,y}(X_l) - r(g_y(X_l) - g_{0,y}(X_l)) \\ - & \qquad (P_{1l} - g_{0,P_1}(X_l) - r(g_{P_1}(X_l) - g_{0,P_1}(X_l)))\theta_1 \\ - & \qquad (P_{2l} - g_{0,P_2}(X_l) - r(g_{P_2}(X_l) - g_{0,P_2}(X_l)))\theta_2]\kappa_{j,l} \end{aligned}$$

$$\partial_r E[I']|_{r=0} = -I_1 - I_2' + I_3' + I_4' - I_5$$

$$\begin{aligned} I_1 &= E\left[(g_{P_1}(X_j) - g_{0,P_1}(X_j))(y_l - g_{0,y}(X_l) - (P_{1l} - g_{0,P_1}(X_l))\theta_{0,1} - (P_{2l} - g_{0,P_2}(X_l))\theta_{0,2})\kappa_{j,l}\right] \\ &= 0 \\ I_2' &= E\left[(P_{1j} - g_{0,P_1}(X_j) - \frac{E[(P_{1m} - g_{0,P_1}(X_m))\kappa_{m,l}|X_l]}{E[\kappa_{m,l}|X_l]})(g_y(X_l) - g_{0,y}(X_l)\kappa_{j,l}\right] \\ &= E\left(E\left[\left(P_{1j} - g_{0,P_1}(X_j) - \frac{E[(P_{1m} - g_{0,P_1}(X_m))\kappa_{m,l}|X_l]}{E[\kappa_{m,l}|X_l]}\right)\kappa_{j,l}|X_l\right](g_y(X_l) - g_{0,y}(X_l)\right) \\ &= E\left(E\left[(P_{1j} - g_{0,P_1}(X_j))\,\kappa_{j,l} - \frac{E[(P_{1m} - g_{0,P_1}(X_m))\kappa_{m,l}|X_l]}{E[\kappa_{m,l}|X_l]}\kappa_{j,l}|X_l\right](g_y(X_l) - g_{0,y}(X_l)\right) \\ &= 0 \end{aligned}$$

because $E\left[(P_{1j} - g_{0,P_1}(X_j))\,\kappa_{j,l} - \frac{E[(P_{1m} - g_{0,P_1}(X_m))\kappa_{m,l}|X_l]}{E[\kappa_{m,l}|X_l]}\kappa_{j,l}|X_l\right] = 0$.

$$\begin{aligned} I_3' &= E\left[(P_{1j} - g_{0,P_1}(X_j) - \frac{E[(P_{1m} - g_{0,P_1}(X_m))\kappa_{m,l}|X_l]}{E[\kappa_{m,l}|X_l]})(g_{P_1}(X_l) - g_{0,P_1}(X_l))\theta_1\kappa_{j,l}\right] = 0 \\ I_4' &= E\left[(P_{1j} - g_{0,P_1}(X_j) - \frac{E[(P_{1m} - g_{0,P_1}(X_m))\kappa_{m,l}|X_l]}{E[\kappa_{m,l}|X_l]})(g_{P_2}(X_l) - g_{0,P_2}(X_l))\theta_2\kappa_{j,l}\right] = 0 \\ I_5 &= E\left[\frac{\left(g_{\tilde{P}_{1,j}}(X_l) - g_{0,\tilde{P}_{1,j}}(X_l)\right)g_{0,\kappa_{m,l}}(X_l) - \left(g_{\kappa_{j,l}}(X_l) - g_{0,\kappa_{m,l}}(X_l)\right)g_{0,\tilde{P}_{1,j}}(X_l)}{(g_{0,\kappa_{m,l}}(X_l))^2}\epsilon_l\kappa_{j,l}\right] = 0 \end{aligned}$$

Since $E(\epsilon_l|X_l,Z_l) = 0$.

Hence, $\partial_g E[\Psi(D;\theta_0,g_0)] = 0$.

□

2.4 Proof of Proposition 4:

Proof.

$$\begin{aligned}\tilde{\theta}_{n,o} = \quad & \left[\tfrac{1}{n(n-1)}\textstyle\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\tilde{P}_j - \tfrac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)\tilde{P}_l'\right]^{-1} \\ & \left[\tfrac{1}{n(n-1)}\textstyle\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\tilde{P}_j - \tfrac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)(\tilde{P}_l\theta_0+\epsilon_l)\right] \\ = \quad & \theta_0 + \left[\tfrac{1}{n(n-1)}\textstyle\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\tilde{P}_j - \tfrac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)\tilde{P}_l'\right]^{-1} \\ & \left[\tfrac{1}{n(n-1)}\textstyle\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\tilde{P}_j - \tfrac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)\epsilon_l\right]\end{aligned}$$

Denote $A_n = \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)\tilde{P}_l'$ and

$B_n = \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)\epsilon_l$. We first show that A_n is a U-statistic and find its probability limit. Then we show that B_n is also a U-statistic and find its probability limit.

To show that A_n is a U-statistic, notice that

$$A_n = \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)\tilde{P}_l'$$

$$= \frac{1}{2}\frac{2}{n(n-1)}\sum_{j<l}^{n}\left(\kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)\tilde{P}_l' + \kappa_{l,j}\left(\tilde{P}_l - \frac{g_{0,\tilde{P}_m}(X_j)}{g_{0,\kappa_{m,j}}(X_j)}\right)\tilde{P}_j'\right)$$

According to WLLN for U-statistics under Assumption 4,

$$A_n \xrightarrow{p} A \qquad \text{with} \qquad A \equiv E\left[\kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)\tilde{P}_l'\right]$$

and under Assumption 4(*iii*), A is nonsingular.

To show that B_n is also a U-statistic, notice that:

$$B_n = \frac{1}{n(n-1)}\sum_{j<l}^{n}\left(\kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)\epsilon_l + \kappa_{l,j}\left(\tilde{P}_l - \frac{g_{0,\tilde{P}_m}(X_j)}{g_{0,\kappa_{m,j}}(X_j)}\right)\epsilon_j\right)$$

Define $h(\tilde{p}_1, e_1, z_1, x_1; \tilde{p}_2, e_2, z_2, x_2) = \kappa_{1,2}\tilde{p}_1 e_2 + \kappa_{2,1}\tilde{p}_2 e_1 - \kappa_{1,2}\frac{g_{0,\tilde{P}_1}(x_2)}{g_{0,\kappa_{1,2}}(x_2)}e_2 - \kappa_{2,1}\frac{g_{0,\tilde{P}_2}(x_1)}{g_{0,\kappa_{2,1}}(x_1)}e_1$. Since h is a symmetric function of observations 1 and 2, a U-statistic with kernel h is defined as

$$B_n' = \frac{2}{n(n-1)}\sum_{j<l}^{n}h(\tilde{P}_j, \epsilon_j, Z_j, X_j; \tilde{P}_l, \epsilon_l, Z_l, X_l) \qquad \text{and} \qquad B_n = \frac{1}{2}B_n'$$

And we have:

$$\begin{aligned}
E(B_n') &= E(\kappa_{j,l}\tilde{P}_j\epsilon_l + \kappa_{l,j}\tilde{P}_l\epsilon_j) - \left(E\left(\kappa_{j,l}\frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\epsilon_l\right) + E\left(\kappa_{l,j}\frac{g_{0,\tilde{P}_m}(X_j)}{g_{0,\kappa_{m,j}}(X_j)}\epsilon_j\right)\right) \\
&= 2E(\kappa_{j,l}\tilde{P}_j\epsilon_l) - 2E\left(\kappa_{j,l}\frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\epsilon_l\right) \\
&= 2\int_{\mathbb{R}^{q_z}} E[\tilde{P}_j\epsilon_l e^{it'(Z_j-Z_l)}]d\mu(t) - 2\int_{\mathbb{R}^{q_z}} E[\frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\epsilon_l e^{it'(Z_j-Z_l)}]d\mu(t) \\
&= 2\int_{\mathbb{R}^{q_z}} E[\tilde{P}_j e^{it'Z_j}\epsilon_l e^{-it'Z_l}]d\mu(t) - 2\int_{\mathbb{R}^{q_z}} E[\frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)} e^{it'Z_j}\epsilon_l e^{-it'Z_l}]d\mu(t) \\
&= 2\int_{\mathbb{R}^{q_z}} E[\tilde{P}_j e^{it'Z_j}]E[\epsilon_l e^{-it'Z_l}]d\mu(t) - 2\int_{\mathbb{R}^{q_z}} E[e^{it'Z_j}]E[\frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\epsilon_l e^{-it'Z_l}]d\mu(t) \\
&= 0 \qquad \text{since } E[\epsilon_l' e^{-it'Z_l}] = 0 \text{ and } E[\epsilon_l|X_l, Z_l] = 0.
\end{aligned}$$

Hence, $E(B_n) = 0$. According to WLLN for U statistics, we have $B_n \xrightarrow{p} 0$, and we conclude that $\tilde{\beta}_n$ is a consistent estimator of β_0.

To derive the asymptotic normality, we first need to compute the asymptotic variance for the U-statistic B_n', which means that we need to find the variance for $E(h(\tilde{P}_1, \epsilon_1, Z_1, X_1; \tilde{P}_2, \epsilon_2, Z_2, X_2)|\tilde{P}_1 = \tilde{p}_1, \epsilon_1 = e_1, Z_1 = z_1, X_1 = x_1)$.
Let $h_1(\tilde{p}_1, e_1, z_1, x_1) \equiv E(h(\tilde{P}_1, \epsilon_1, Z_1, X_1; \tilde{P}_2, \epsilon_2, Z_2, X_2)|\tilde{P}_1 = \tilde{p}_1, \epsilon_1 = e_1, Z_1 = z_1, X_1 = x_1)$. We have:

$$h(\tilde{p}_1, e_1, z_1, x_1; \tilde{p}_2, e_2, z_2, x_2) = \kappa_{1,2}\tilde{p}_1 e_2 + \kappa_{2,1}\tilde{p}_2 e_1 - \kappa_{1,2}\frac{g_{0,\tilde{P}_1}(x_2)}{g_{0,\kappa_{1,2}}(x_2)}e_2 - \kappa_{2,1}\frac{g_{0,\tilde{P}_2}(x_1)}{g_{0,\kappa_{2,1}}(x_1)}e_1$$

and

$$\begin{aligned}
h_1(\tilde{p}_1, e_1, z_1, x_1) &= E[\int_{\mathbb{R}^{q_z}} e^{it'(z_1-Z_2)}d\mu(t)\tilde{p}_1\epsilon_2 + \int_{R^{q_z}} e^{it'(Z_2-z_1)}d\mu(t)\tilde{P}_2 e_1 \\
&- \int_{\mathbb{R}^{q_z}} e^{it'(z_1-Z_2)}d\mu(t)\frac{g_{0,\tilde{P}_1}(X_2)}{g_{0,\kappa_{1,2}}(X_2)}\epsilon_2 - \int_{R^{q_z}} e^{it'(Z_2-z_1)}d\mu(t)\frac{g_{0,\tilde{P}_2}(x_1)}{g_{0,\kappa_{2,1}}(x_1)}e_1] \\
&= E[\int_{\mathbb{R}^{q_z}} e^{it'(z_1-Z_2)}d\mu(t)\tilde{p}_1\epsilon_2] + E[\int_{R^{q_z}} e^{it'(Z_2-z_1)}d\mu(t)\tilde{P}_2 e_1] \\
&- E[\int_{\mathbb{R}^{q_z}} e^{it'(z_1-Z_2)}d\mu(t)\frac{g_{0,\tilde{P}_1}(X_2)}{g_{0,\kappa_{1,2}}(X_2)}\epsilon_2] - E[\int_{R^{q_z}} e^{it'(Z_2-z_1)}d\mu(t)\frac{g_{0,\tilde{P}_2}(x_1)}{g_{0,\kappa_{2,1}}(x_1)}e_1]
\end{aligned}$$

The first element of the right hand side is

$$\begin{aligned}
E[\int_{\mathbb{R}^{q_z}} e^{it'(z_1-Z_2)}d\mu(t)\tilde{p}_1\epsilon_2] &= \int_{\mathbb{R}^{q_z}} E[e^{it'z_1}e^{-it'Z_2}\tilde{p}_1\epsilon_2]d\mu(t) \\
&= \int_{\mathbb{R}^{q_z}} e^{it'z_1}\tilde{p}_1 E[e^{-it'Z_2}\epsilon_2]d\mu(t) = 0
\end{aligned}$$

The third term is

$$\begin{aligned}
E[\int_{\mathbb{R}^{q_z}} e^{it'(z_1-Z_2)}d\mu(t)\frac{g_{0,\tilde{P}_1}(X_2)}{g_{0,\kappa_{1,2}}(X_2)}\epsilon_2] &= \int_{\mathbb{R}^{q_z}} E[e^{it'z_1}e^{-it'Z_2}\frac{g_{0,\tilde{P}_1}(X_2)}{g_{0,\kappa_{1,2}}(X_2)}\epsilon_2]d\mu(t) \\
&= \int_{\mathbb{R}^{q_z}} e^{it'z_1}\tilde{p}_1 E[e^{-it'Z_2}\frac{g_{0,\tilde{P}_1}(X_2)}{g_{0,\kappa_{1,2}}(X_2)}\epsilon_2]d\mu(t) = 0
\end{aligned}$$

The second term of the right hand side is

$$\begin{aligned}
E[\int_{R^{q_z}} e^{it'(Z_2-z_1)}d\mu(t)\tilde{P}_2 e_1] &= \int_{\mathbb{R}^{q_z}} E[e^{it'Z_2}e^{-it'z_1}\tilde{P}_2 e_1]d\mu(t) \\
&= \int_{\mathbb{R}^{q_z}} e^{-it'z_1}e_1 E[e^{it'Z_2}\tilde{P}_2]d\mu(t)
\end{aligned}$$

The fourth term is

$$\begin{aligned}
E[\int_{R^{q_z}} e^{it'(Z_2-z_1)}d\mu(t)\frac{g_{0,\tilde{P}_2}(x_1)}{g_{0,\kappa_{2,1}}(x_1)}e_1] &= \int_{\mathbb{R}^{q_z}} E[e^{it'Z_2}e^{-it'z_1}\frac{g_{0,\tilde{P}_2}(x_1)}{g_{0,\kappa_{2,1}}(x_1)}e_1]d\mu(t) \\
&= \int_{\mathbb{R}^{q_z}} e^{-it'z_1}e_1\frac{g_{0,\tilde{P}_2}(x_1)}{g_{0,\kappa_{2,1}}(x_1)}E[e^{it'Z_2}]d\mu(t)
\end{aligned}$$

Hence, $h_1(\tilde{p}_1, e_1, z_1, x_1) = \int_{\mathbb{R}^{q_z}} e^{-it'z_1}e_1 E[e^{it'Z_2}\tilde{P}_2]d\mu(t) - \int_{\mathbb{R}^{q_z}} e^{-it'z_1}e_1\frac{g_{0,\tilde{P}_2}(x_1)}{g_{0,\kappa_{2,1}}(x_1)}E[e^{it'Z_2}]d\mu(t)$.
$h_1(\tilde{p}_1, e_1, z_1, x_1) = \int_{\mathbb{R}^{q_z}} e^{-it'z_1}e_1\left(E[e^{it'Z_2}\tilde{P}_2] - \frac{g_{0,\tilde{P}_2}(x_1)}{g_{0,\kappa_{2,1}}(x_1)}E[e^{it'Z_2}]\right)d\mu(t)$
Since $E(h_1(\tilde{P}_1, \epsilon_1, Z_1, X_1)) = 0$, we have:

$$\begin{aligned}
Var[h_1(\tilde{P}_1, \epsilon_1, Z_1, X_1)] &= Var\left[\int_{\mathbb{R}^{q_z}} e^{-it'Z_1}\epsilon_1\left(E[e^{it'Z_2}\tilde{P}_2] - \frac{g_{0,\tilde{P}_2}(X_1)}{g_{0,\kappa_{2,1}}(X_1)}E[e^{it'Z_2}]\right)d\mu(t)\right] \\
&= Var\left[\int_{\mathbb{R}^{q_z}} e^{-it'Z_j}\epsilon_j\left(E[e^{it'Z_l}\tilde{P}_l] - \frac{g_{0,\tilde{P}_m}(X_j)}{g_{0,\kappa_{m,j}}(X_j)}E[e^{it'Z_l}]\right)d\mu(t)\right]
\end{aligned}$$

Following Hoeffding (1948b), the asymptotic distribution for U-statistics yields:

$$\sqrt{n}(B_n' - 0) \xrightarrow{d} \mathcal{N}(0, 4Var[h_1(\tilde{P}_1, \epsilon_1, Z_1, X_1)])$$

Thus,

$$\sqrt{n}(\tilde{\theta}_{n,o} - \theta_0) \xrightarrow{d} N\left(0, A^{-1}Var[h_1(\tilde{P}_1, \epsilon_1, Z_1, X_1)]\left(A^{-1}\right)'\right)$$

with $A = E\left[\kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)\tilde{P}_l'\right]$. □

2.5 Proof of Theorem 3.3

Proof. Recall that

$$\widetilde{\widehat{y}}_i = y_i - \hat{g}_y(X_i) = y_i - E(y_i|X_i) + E(y_j|X_i) - \hat{g}_y(X_i)$$

For the first two terms of the right hand side, we have

$$\begin{aligned} y_i - E(y_i|X_i) &= (P_i - E(P_i|X_i))'\theta_0 + \epsilon_i \\ &= (P_i - \hat{g}_P(X_i))'\theta_0 + (\hat{g}_P(X_i) - E(P_i|X_i))'\theta_0 + \epsilon_i \end{aligned}$$

In matrix form, the feasible estimator writes:

$$\hat{\theta}_{n,o} = \left[\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\hat{\tilde{P}}_j - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right)\widehat{\tilde{P}_l}'\right]^{-1}\left[\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\hat{\tilde{P}}_j - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right)\hat{\tilde{y}}_l\right]$$

Define $C_n = \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\hat{\tilde{P}}_j - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right)\widehat{\tilde{P}_l}'$. We get:

$$\begin{aligned} \hat{\theta}_{n,o} &= [C_n]^{-1}[\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\hat{\tilde{P}}_j - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right)[y_l - E(y_l|X_l) + E(y_l|X_l) - \hat{g}_y(X_l)]] \\ &= [C_n]^{-1}[\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\hat{\tilde{P}}_j - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right)[(P_l - \hat{g}_P(X_l))'\theta_0 \\ &\quad +(\hat{g}_P(X_l) - E(P_l|X_l))'\theta_0 + \epsilon_l + E(y_l|X_l) - \hat{g}_y(X_l)]] \\ &= \theta_0 + [C_n]^{-1}[\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\hat{\tilde{P}}_j - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right) \\ &\quad [(\hat{g}_P(X_l) - E(P_l|X_l))'\theta_0 + \epsilon_l + E(y_l|X_l) - \hat{g}_y(X_l)]] \end{aligned}$$

Consider now,

$$\begin{aligned} A_n - C_n &= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)\tilde{P}_l' \\ &\quad -\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\hat{\tilde{P}}_j - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right)\widehat{\tilde{P}_l}' \\ &= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)\tilde{P}_l' \\ &\quad -\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\tilde{P}_j + \hat{\tilde{P}}_j - \tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)} + \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)} - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right)\widehat{\tilde{P}_l}' \\ &= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}[\left(\hat{\tilde{P}}_j - \tilde{P}_j + \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)} - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right)\widehat{\tilde{P}_l}' \\ &\quad + \left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)(-\hat{\tilde{P}}_l + \tilde{P}_l)] \\ &\overset{P}{\to} 0 \end{aligned}$$

Hence, we have $\texttt{Plim}C_n = A$, since we showed in the proof of Proposition 4 that $\texttt{Plim}A_n = A$.

Define now the following quantities:

$$
\begin{aligned}
D_n &= \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{l \neq j}^{n} \kappa_{j,l} \left(\widehat{\tilde{P}}_j - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)} \right) \left[(\hat{g}_P(X_l) - E(P_l|X_l))'\theta_0 + \epsilon_l + E(y_l|X_l) - \hat{g}_y(X_l) \right] \\
E_n &= \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{l \neq j}^{n} \kappa_{j,l} \left(\widehat{\tilde{P}}_j - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)} \right) (\hat{g}_P(X_l) - E(P_l|X_l))'\theta_0 \\
F_n &= \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{l \neq j}^{n} \kappa_{j,l} \left(\widehat{\tilde{P}}_j - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)} \right) \epsilon_l \\
G_n &= \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{l \neq j}^{n} \kappa_{j,l} \left(\widehat{\tilde{P}}_j - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)} \right) \left[E(y_l|X_l) - \hat{g}_y(X_l) \right]
\end{aligned}
$$

We have, $D_n = E_n + F_n + G_n$. For consistency, we show that the probability limits for E_n, F_n, and G_n are all zero.

$$\begin{aligned}
E_n &= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\hat{\tilde{P}}_j - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right)(\hat{g}_P(X_l) - E(P_l|X_l))'\theta_0 \\
&= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\tilde{P}_j + \hat{\tilde{P}}_j - \tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)} + \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)} - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right) \\
&\quad (\hat{g}_P(X_l) - E(P_l|X_l))'\theta_0 \\
&= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)} + \hat{\tilde{P}}_j - \tilde{P}_j + \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)} - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right) \\
&\quad (\hat{g}_P(X_l) - E(P_l|X_l))'\theta_0 \\
&= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)(\hat{g}_P(X_l) - E(P_l|X_l))'\theta_0 \\
&+ \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\hat{\tilde{P}}_j - \tilde{P}_j + \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)} - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right)(\hat{g}_P(X_l) - E(P_l|X_l))'\theta_0 \\
&\xrightarrow{P} 0 \\
F_n &= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)} + \hat{\tilde{P}}_j - \tilde{P}_j + \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)} - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right)\epsilon_l \\
&= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)\epsilon_l \\
&+ \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\hat{\tilde{P}}_j - \tilde{P}_j + \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)} - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right)\epsilon_l \\
&\xrightarrow{P} 0 \\
G_n &= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)[E(y_l|X_l) - \hat{g}_y(X_l)] \\
&+ \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\hat{\tilde{P}}_j - \tilde{P}_j + \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)} - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right)[E(y_l|X_l) - \hat{g}_y(X_l)] \\
&\xrightarrow{P} 0
\end{aligned}$$

All in all, we have $\texttt{Plim}D_n = \texttt{Plim}(E_n + F_n + G_n) = 0$, so $\hat{\theta}_{n,o} \xrightarrow{P} \theta_0$.

In addition, we have:

$$\sqrt{n}(\hat{\theta}_{n,o} - \theta_0) = [C_n]^{-1}\sqrt{n}[E_n + F_n + G_n]$$

And we study each term separately:

$$\begin{aligned}
\sqrt{n}F_n &= \sqrt{n}B_n + \sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\widehat{\tilde{P}}_j - \tilde{P}_j + \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)} - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right)\epsilon_l \\
\sqrt{n}E_n &= \sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)(\hat{g}_P(X_l) - E(P_l|X_l))'\theta_0 \\
&+ \sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\widehat{\tilde{P}}_j - \tilde{P}_j + \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)} - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right)(\hat{g}_P(X_l) - E(P_l|X_l))'\theta_0
\end{aligned}$$

$$\begin{aligned}
\sqrt{n}G_n &= \sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\tilde{P}_j - \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)}\right)[E(y_l|X_l) - \hat{g}_y(X_l)] \\
&+ \sqrt{n}\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\left(\widehat{\tilde{P}}_j - \tilde{P}_j + \frac{g_{0,\tilde{P}_m}(X_l)}{g_{0,\kappa_{m,l}}(X_l)} - \frac{\widehat{g_{0,\tilde{P}_m}}(X_l)}{\widehat{g_{0,\kappa_{m,l}}}(X_l)}\right)[E(y_l|X_l) - \hat{g}_y(X_l)]
\end{aligned}$$

From Assumption 5, $\sqrt{n}[E_n + F_n + G_n] = \sqrt{n}B_n + o_p(1)$.

Hence,

$$\sqrt{n}(\hat{\theta}_{n,o} - \theta_0) = [C_n]^{-1}\sqrt{n}B_n + o_p(1)$$

$\hat{\theta}_{n,o}$ and $\tilde{\theta}_{n,o}$ share the same asymptotic distribution. □

2.6 Consistent Estimator of the Asymptotic Variance (3.11)

$$\begin{aligned}
& Var\left[\int_{\mathbb{R}^{q_z}} e^{-it'Z_j}\epsilon_j\left(E[e^{it'Z_l}\tilde{P}_l] - \frac{g_{0,\tilde{P}_m}(X_j)}{g_{0,\kappa_{m,j}}(X_j)}E[e^{it'Z_l}]\right)d\mu(t)\right] \\
= \; & Var\left[E_l\left[\int_{\mathbb{R}^{q_z}} e^{it'(Z_l - Z_j)}\epsilon_j\left(\tilde{P}_l - \frac{g_{0,\tilde{P}_m}(X_j)}{g_{0,\kappa_{m,j}}(X_j)}\right)d\mu(t)\right]\right] \\
= \; & Var\left[E_l\left[k(Z_l - Z_j)\left(\tilde{P}_l - \frac{g_{0,\tilde{P}_m}(X_j)}{g_{0,\kappa_{m,j}}(X_j)}\right)\epsilon_j\right]\right] \\
= \; & E\left[\left(E_l\left[k(Z_l - Z_j)\left(\tilde{P}_l - \frac{g_{0,\tilde{P}_m}(X_j)}{g_{0,\kappa_{m,j}}(X_j)}\right)\right]\right)\left(E_l\left[k(Z_l - Z_j)\left(\tilde{P}_l - \frac{g_{0,\tilde{P}_m}(X_j)}{g_{0,\kappa_{m,j}}(X_j)}\right)'\right]\right)[\epsilon_j]^2\right]
\end{aligned}$$

Under heteroskedasticity, the estimator for variance of the feasible-RSMD is

$$[n(n-1)C_n]^{-1}\sum_{j=1}^{n}((\sum_{l=1}^{n}\kappa_{j,l}\left(\widehat{\tilde{P}}_l - \frac{\widehat{g_{0,\tilde{P}_l}}(X_j)}{\widehat{g_{0,\kappa_{m,j}}}(X_j)}\right))(\sum_{l=1}^{n}\kappa_{j,l}\left(\widehat{\tilde{P}}_l - \frac{\widehat{g_{0,\tilde{P}_l}}(X_j)}{\widehat{g_{0,\kappa_{m,j}}}(X_j)}\right)')\hat{\epsilon}_j^2)[n(n-1)C_n']^{-1}$$

3 R-SMD Estimator and its Properties from Antoine and Sun (2021)

Under Assumption 4 and FOC defined in Equation (2.8)

$$\theta_0 = [E(\kappa_{j,l}\tilde{P}_j\tilde{P}_l')]^{-1}E(\kappa_{j,l}\tilde{P}_j\tilde{y}_l)$$

$$\tilde{\theta}_{n,u} = \left[\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{P}_j\tilde{P}_l'\right]^{-1}\left[\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\tilde{P}_j\tilde{y}_l\right]$$

Additionally, we show the similar consistency and Asymptotic normality for $\tilde{\theta}_{n,u}$. Under Assumption 4 and iid assumption for the sample, $\tilde{\theta}_{n,u}$ is consistent for θ_0, and the asymptotic distribution is:

$$\sqrt{n}(\tilde{\theta}_{n,u}-\theta_0)\xrightarrow{d}N\left(0,E\left[\kappa_{j,l}\tilde{P}_j\tilde{P}_l'\right]^{-1}Var[h_1'(\epsilon_1,Z_1,X_1)]\left(E\left[\kappa_{j,l}\tilde{P}_j\tilde{P}_l'\right]^{-1}\right)'\right)$$

where $h_1'(e_1,z_1,x_1) = E[\int_{\mathbb{R}^{q_z}} e^{it'(z_1-Z_2)}d\mu(t)\tilde{p}_1\epsilon_2 + \int_{R^{q_z}} e^{it'(Z_2-z_1)}d\mu(t)\tilde{P}_2e_1$.

Replace every nuisance parameter with its estimate, we have the feasible estimator $\hat{\theta}_{n,u}$, the R-SMD estimator is in the following.

$$\hat{\theta}_{n,u} = \left[\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\hat{\tilde{P}}_j\widehat{\tilde{P}_l}'\right]^{-1}\left[\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\kappa_{j,l}\hat{\tilde{P}}_j\hat{\tilde{y}}_l\right]$$

since $E[\kappa_{j,l}\hat{\tilde{P}}_j\widehat{\tilde{P}_l}']$ is invertible.

Under Assumptions 4 - 5, $\hat{\theta}_{n,u}$ is consistent for θ_0, and the asymptotic distribution is:

$$\sqrt{n}(\hat{\theta}_{n,u}-\theta_0)\xrightarrow{d}N\left(0,E\left[\kappa_{j,l}\tilde{P}_j\tilde{P}_l'\right]^{-1}Var[h_1'(\epsilon_1,Z_1,X_1)]\left(E\left[\kappa_{j,l}\tilde{P}_j\tilde{P}_l'\right]^{-1}\right)'\right)$$

where $h_1'(e_1,z_1,x_1) = E[\int_{\mathbb{R}^{q_z}} e^{it'(z_1-Z_2)}d\mu(t)\tilde{p}_1\epsilon_2 + \int_{R^{q_z}} e^{it'(Z_2-z_1)}d\mu(t)\tilde{P}_2e_1$.

Supplementary Appendix

S1. Robustness Check: When the Model is not Sparse

All of the other simulation results are based on the sparsity assumption. In this section, Table B.5 reports the results when the sparsity assumption is unsatisfied. We choose the DGP where all of the β_{1q} and β_{2q} are not 0. When $q \leq 5$, $\beta_{1q} = 1$ and $\beta_{2q} = 3$. When $q > 5$, $\beta_{1q} = \beta_{2q} = 0.05$. The D-RSMD with the Lasso method as the nuisance parameter estimator is affected by the violation of the assumption. It is no surprise since we need the sparsity assumption to conduct the Lasso method. The magnitude of the Med.Bias, MAD and Med.SE are not changing. RRs are higher.

Instrument with 3 values ($P = 30$)							
		Estimator	*Ins*	Med.Bias	MAD	Med.SE	RR
$n = 3000$	θ_{w0}	D-RSMD	Z_2	-0.045	0.123	0.158	0.063
		D-RSMD	(Z_2, Z_2X_1)	-0.019	0.102	0.119	0.057
		GMM (Oracle)	(Z_2, Z_2X_1)	0.000	0.089	0.130	0.056
$n = 3000$	θ_{wx0}	D-RSMD	Z_2	-0.032	0.162	0.238	0.076
		D-RSMD	(Z_2, Z_2X_1)	-0.053	0.069	0.084	0.094
		GMM (Oracle)	(Z_2, Z_2X_1)	0.000	0.048	0.071	0.052

Table B.5: Instruments with 3 values

Note: Simulation Results for θ_{w0} and θ_{wx0} in the benchmark model using D-RSMD estimator 5000 replications. We report the Monte-Carlo Median Bias (Med.Bias), Median Absolute Deviation (MAD), median of asymptotic standard error under heteroskedasticity (Med.SE), and the Rejection Rate (RR) using a 5% t-test.

S2. Potential Outcome Framework

We have two potential outcomes under traditional potential outcome framework. They are y_0 and y_1.

$y_0 = \mu_0 + f_0(X) + e_0,$

$y_1 = \mu_1 + f_1(X) + e_1.$

The observed outcome y is

$y = (1 - W) \cdot (\mu_0 + f_0(X) + e_0) + W \cdot (\mu_1 + f_1(X) + e_1).$

$y = \mu_0 + W(\mu_1 - \mu_0) + f_0(X) + e_0 + W \cdot (f_1(X) + e_1 - f_0(X) - e_0)$

That is,

$y = \mu_0 + W(\mu_1 - \mu_0) + f_0(X) + W \cdot (f_1(X) - f_0(X)) + e_0 + W(e_1 - e_0)$

Define $\epsilon = e_0 + W(e_1 - e_0)$

$y = \mu_0 + W(\mu_1 - \mu_0) + W \cdot (f_1(X) - f_0(X)) + f_0(X) + \epsilon$

If we assume difference between $f_0(X)$ and $f_1(X)$ is linear in θ_{wx0}, the equation becomes

$y = \mu_0 + \theta_{w0}W + W \cdot X_1'\theta_{wx0} + f_0(X) + \epsilon.$

This assumption is that $f_0(X) - f_1(X)$ is $\cdot X_1'\theta_{wx0}$. Recall that we could include nonlinear terms of X_1 inside the interaction terms in a more general model. The assumption suggests that after taking into consideration many complex functions of X, the difference is mainly caused by X_1. It is reasonable. There are other works that make the same assumption, for instance, Nekipelov et al. (2018).

Assuming that all of the participants are compliers, their response to treatment is heterogeneous due to the interaction term $W \quad \cdot X'_{i1}\theta_{wx0}$. This problem also belongs to the essential heterogeneous problem or random coefficient problem.

S3. Illustration of Identification Issue for a Specific Model

Prove the identification for the key parameters using the RSMD method.

Proof. Consider the simplest model in the following.

$$\begin{aligned}
y_i &= \beta_1 X_{i,1} + \beta_2 X_{i,2} + u_i \\
X_{i,1} &= \pi_1 W_i + v_{i,1} \\
X_{i,2} &= \pi_2 W_i + v_{i,2}
\end{aligned}$$

We need to prove that $E(\kappa_{j,l} X_j X'_l)$ is nonsingular. Because the new estimator becomes

$$\begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix} = [E(\kappa_{j,l} X_j X'_l)]^{-1} E(\kappa_{j,l} X_j y_l) = \left[E\left(\kappa_{j,l}\begin{pmatrix}X_{j,1} \\ X_{j,2}\end{pmatrix}\begin{pmatrix}X_{l,1} \\ X_{l,2}\end{pmatrix}'\right)\right]^{-1} E(\kappa_{j,l} X_j y_l)$$

To show that $E(\kappa_{j,l} X_j X'_l)$ is nonsingular, we consider the associated quadratic form, and show that it is positive definite. For any a real vector of size p, we have:

$$\begin{aligned}
E(a' X_j X'_l a \kappa_{j,l}) &= E(\kappa_{j,l} a' E(X_j|W_j) E(X'_l a|W_l)) \\
&= E(\int_{R^{q_w}} e^{it'(W_j - W_l)} d\mu(t) a' E(X_j|W_j) E(X'_l|W_l) a) \\
&= \int_{R^{q_w}} E[e^{it'(W_j - W_l)} a' E(X_j|W_j) E(X'_l|W_l) a] d\mu(t) \\
&= \int_{R^{q_w}} E[e^{it'W_j} a' E(X_j|W_j) E(X'_l|W_l) a e^{-it'W_l}] d\mu(t) \\
&= \int_{R^{q_w}} E[a' e^{it'W_j} E(X_j|W_j)] E[E(X'_l|W_l) a e^{-it'W_l}] d\mu(t) \\
&= \int_{R^{q_w}} E[a' e^{it'W_j} E(X_j|W_j)] E[E(X'_j|W_j) a e^{-it'W_j}] d\mu(t) \\
&= \int_{R^{q_w}} \left| \left(\int_{R^{q_w}} a' e^{it'W_j} E(X_j|W_j) f_W(W_j) dW_j \right) \right|^2 d\mu(t) \\
&= (2\pi)^{2q_w} \int_{R^{q_w}} \left| \left(\mathcal{F}[a' E(X_j|W_j) f_W(W_j)](t) \right) \right|^2 d\mu(t) \\
&\geq 0
\end{aligned}$$

with μ strictly positive on $\mathbb{R}^p$ and $\mathcal{F}[g]$ the Fourier transform of a well-defined function $g(.)$ on $\mathbb{R}^{q_w}$ formally defined as,

$$\mathcal{F}[g](t) = \frac{1}{(2\pi)^{q_w}} \int \exp^{it'u} g(u) du \,. \tag{3.1}$$

We then have:

$$\begin{aligned} a'E(\kappa_{j,l}X_jX_l')a = 0 \quad &\Leftrightarrow \quad \exists\ a \neq 0\ s.t.\ a'E(X_j|W_j)f(W_j) = 0\ a.s. \\ &\Leftrightarrow \quad \exists\ a \neq 0\ s.t.\ a'E(X_j|W_j) = 0\ a.s. \end{aligned}$$

In the specific case,

$$\Leftrightarrow \exists\ a \neq 0\ s.t.\ a'E(X_j|W_j) = a_1\pi_1W_i + a_2\pi_2W_i = (a_1\pi_1 + a_2\pi_2)W_i = 0\ a.s.$$

β_1 and β_2 are not identified. To identify β_1 and β_2, the model need to have a nonlinear part of W_i or another control variable Z_i inside one of the equations for X_i. □

Appendix C

Estimation of Heterogeneous Treatment Effects: Oregon Health Insurance Experiment

1 Results of Oregon Health Insurance Experiment

Panel A: Heterogeneous treatment effects										
Debt for Health	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for θ_{w0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-0.088		-0.122		-0.200***		-0.226***		-0.020
		(0.076)		(0.077)		(0.062)		(0.053)		(0.116)
GMM-Lasso		-0.101		-0.121		-0.205**		-0.318***		-0.251
		(0.141)		(0.122)		(0.088)		(0.095)		(0.159)
DRSMD-Lasso	-0.203***	-0.108	-0.177**	-0.071	-0.146**	-0.167**	-0.241***	-1.131	-0.162	-0.278**
	(0.075)	(0.252)	(0.080)	(0.148)	(0.071)	(0.078)	(0.063)	(0.691)	(0.116)	(0.126)
Estimator for θ_{wx0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-0.002*		-0.001		0.002		0.001		-0.003
		(0.001)		(0.001)		(0.001)		(0.001)		(0.003)
GMM-Lasso		-0.002		-0.001		0.000		-0.001		-0.001
		(0.001)		(0.001)		(0.001)		(0.002)		(0.004)
DRSMD-Lasso	0.000	-0.007	0.001**	-0.004*	0.000	0.003	0.001***	0.045	0.000	0.005**
	(0.000)	(0.006)	(0.000)	(0.002)	(0.000)	(0.002)	(0.000)	(0.034)	(0.000)	(0.002)
N	3504		3596		3941		4599		2932	

Table C.1: Heterogeneous Treatment Effects of Medicaid on Debt for Health

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid × hhincome.

Average Treatment Effects (Local)										
Debt for Health	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for LATE	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-0.216***		-0.196**		-0.078		-0.166**		-0.251*
		(0.072)		(0.079)		(0.069)		(0.066)		(0.144)
GMM-Lasso		-0.248**		-0.195		-0.223*		-0.362***		-0.318
		(0.149)		(0.145)		(0.124)		(0.133)		(0.281)
DRSMD-Lasso	-0.211**	-0.613**	-0.124	-0.427**	-0.156**	0.064	-0.170***	2.293	-0.122	0.189
	(0.090)	(0.276)	(0.084)	(0.174)	(0.069)	(0.185)	(0.062)	(1.886)	(0.103)	(0.163)
N	3504		3596		3941		4599		2932	

Table C.2: Heterogeneous Treatment Effects of Medicaid on Debt for Health

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid × hhincome. The expression for LATE is $\theta_{w0} + \theta_{wx0}E(X)$.

Panel A: Heterogeneous treatment effects										
Debt for Health	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for θ_{w0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-0.048		-0.125*		-0.210***		-0.190***		-0.244***
		(0.072)		(0.069)		(0.053)		(0.046)		(0.077)
GMM-Lasso		-0.130		-0.087		-0.271***		-0.274***		-0.489***
		(0.121)		(0.122)		(0.101)		(0.091)		(0.156)
DRSMD-Lasso	-0.251**	-0.064	-0.156**	-0.085	-0.155**	-0.218***	-0.269***	-0.215***	-0.141	-0.237**
	(0.098)	(0.086)	(0.087)	(0.086)	(0.076)	(0.066)	(0.071)	(0.055)	(0.112)	(0.102)
Estimator for θ_{wx0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-0.330**		-0.109		0.223**		0.004		0.192
		(0.135)		(0.124)		(0.104)		(0.106)		(0.137)
GMM-Lasso		-0.312		-0.112		0.070		-0.158		0.327
		(0.196)		(0.150)		(0.155)		(0.169)		(0.204)
DRSMD-Lasso	0.059	-0.409***	0.025	-0.119	0.039	0.196*	0.137***	-0.013	0.027	0.237
	(0.045)	(0.158)	(0.039)	(0.139)	(0.039)	(0.118)	(0.039)	(0.116)	(0.047)	(0.164)
Panel B: Homogeneous treatment effects										
GMM	-0.188***	-0.138**	-0.175***	-0.156***	-0.120**	-0.154***	-0.191***	-0.190***	-0.151**	-0.184***
	(0.065)	(0.062)	(0.063)	(0.059)	(0.049)	(0.046)	(0.046)	(0.042)	(0.068)	(0.064)
GMM-Lasso	-0.265*	-0.148	-0.150	-0.096	-0.242**	-0.271***	-0.334***	-0.269***	-0.315*	-0.481***
	(0.140)	(0.120)	(0.142)	(0.121)	(0.120)	(0.101)	(0.114)	(0.090)	(0.183)	(0.156)
DRSMD-Lasso	-0.228***	-0.246***	-0.144*	-0.152*	-0.140**	-0.138**	-0.219***	-0.219***	-0.128	-0.104
	(0.087)	(0.090)	(0.085)	(0.089)	(0.070)	(0.071)	(0.063)	(0.063)	(0.107)	(0.112)
N	3504		3596		3941		4599		2932	

Table C.3: Heterogeneous Treatment Effects of Medicaid on Debt for Health

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid × Above 50% Federal Poverty Line.

Average Treatment Effects										
Debt for Health	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for LATE	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-0.237***		-0.193**		-0.082		-0.192***		-0.116
		(0.076)		(0.073)		(0.058)		(0.060)		(0.080)
GMM-Lasso		-0.308*		-0.158		-0.231*		-0.365***		-0.271
		(0.160)		(0.148)		(0.134)		(0.138)		(0.200)
DRSMD-Lasso	-0.217***	-0.298***	-0.140*	-0.160*	-0.132*	-0.106	-0.189***	-0.222***	-0.123	-0.078
	(0.083)	(0.102)	(0.085)	(0.094)	(0.068)	(0.082)	(0.059)	(0.079)	(0.107)	(0.123)

Table C.4: Heterogeneous Treatment Effects of Medicaid on Debt for Health

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid × Above 50% Federal Poverty Line. The expression for LATE is $\theta_{w0} + \theta_{wx0}E(X)$.

Panel A: Heterogeneous treatment effects										
Debt for Health	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for θ_{w0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-0.193***		-0.199***		-0.113**		-0.196***		-0.157**
		(0.065)		(0.061)		(0.049)		(0.046)		(0.068)
GMM-Lasso		-0.270*		-0.216		-0.211*		-0.351***		-0.295
		(0.147)		(0.136)		(0.110)		(0.112)		(0.182)
DRSMD-Lasso	-0.215**	-0.217**	-0.137*	-0.010	-0.148**	-0.135**	-0.199***	-0.192***	-0.139	-0.141
	(0.087)	(0.103)	(0.082)	(0.337)	(0.068)	(0.068)	(0.062)	(0.062)	(0.107)	(0.107)
Estimator for θ_{wx0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		0.212		1.959		-0.791		2.059		3.074
		(0.900)		(1.353)		(0.652)		(2.040)		(3.827)
GMM-Lasso		0.983		1.904		-0.785		1.083		0.818
		(1.099)		(1.348)		(0.956)		(1.844)		(1.164)
DRSMD-Lasso	0.307**	0.777	0.042	-9.902	-0.320***	-1.448	0.470	-1.725	-0.048	-0.281
	(0.142)	(4.239)	(0.153)	(19.627)	(0.085)	(1.554)	(0.417)	(1.821)	(0.522)	(0.590)
N	3504		3596		3941		4599		2932	

Table C.5: Heterogeneous Treatment Effects of Medicaid on Debt for Health

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid× TANF.

Average Treatment Effects										
Debt for Health	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for LATE	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-0.184**		-0.114		-0.131***		-0.177***		-0.147**
		(0.074)		(0.083)		(0.050)		(0.050)		(0.068)
GMM-Lasso		-0.230		-0.133		-0.228**		-0.341***		-0.293
		(0.156)		(0.151)		(0.111)		(0.115)		(0.182)
DRSMD-Lasso	-0.203**	-0.185	-0.135	-0.440	-0.155**	-0.168**	-0.194***	-0.207***	-0.140	-0.142
	(0.087)	(0.146)	(0.083)	(0.542)	(0.068)	(0.072)	(0.061)	(0.062)	(0.107)	(0.107)
N	3504		3596		3941		4599		2932	

Table C.6: Heterogeneous Treatment Effects of Medicaid on Debt for Health

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid× TANF. The expression for LATE is $\theta_{w0} + \theta_{wx0}E(X)$.

Panel A: Heterogeneous treatment effects										
Debt for Health	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for θ_{w0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-0.174		-0.152		-0.464***		-0.146		-0.424**
		(0.169)		(0.143)		(0.105)		(0.102)		(0.179)
GMM-Lasso		0.392		0.075		-0.302**		0.030		-0.370
		(0.277)		(0.220)		(0.149)		(0.165)		(0.299)
DRSMD-Lasso	-0.326***	-0.136	-0.229**	-0.116	-0.260***	-0.416***	-0.198***	-0.131	-0.194	-0.403*
	(0.110)	(0.195)	(0.100)	(0.166)	(0.095)	(0.119)	(0.075)	(0.115)	(0.137)	(0.210)
Estimator for θ_{wx0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-0.007		-0.010		0.168***		-0.022		0.118
		(0.070)		(0.065)		(0.050)		(0.048)		(0.073)
GMM-Lasso		-0.280***		-0.112		0.042		-0.183**		0.034
		(0.108)		(0.091)		(0.068)		(0.075)		(0.111)
DRSMD-Lasso	0.049**	-0.035	0.040**	-0.009	0.052**	0.135**	0.001	-0.032	0.023	0.117
	(0.022)	(0.077)	(0.020)	(0.071)	(0.021)	(0.054)	(0.018)	(0.052)	(0.026)	(0.082)
N	3504		3596		3941		4599		2932	

Table C.7: Heterogeneous Treatment Effects of Medicaid on Debt for Health

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid× smoke.

Average Treatment Effects										
Debt for Health	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for LATE	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-0.190***		-0.175***		-0.100**		-0.192***		-0.141**
		(0.066)		(0.064)		(0.051)		(0.047)		(0.068)
GMM-Lasso		-0.267*		-0.185		-0.211*		-0.363***		-0.289
		(0.155)		(0.138)		(0.114)		(0.115)		(0.184)
DRSMD-Lasso	-0.210**	-0.220**	-0.135*	-0.136*	-0.146**	-0.123*	-0.197***	-0.201***	-0.138	-0.123
	(0.087)	(0.087)	(0.082)	(0.082)	(0.067)	(0.069)	(0.061)	(0.062)	(0.107)	(0.104)
N	3504		3596		3941		4599		2932	

Table C.8: Heterogeneous Treatment Effects of Medicaid on Debt for Health

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid× smoke. The expression for LATE is $\theta_{w0} + \theta_{wx0}E(X)$.

Robustness Check: Other Results of Oregon Health Insurance Experiment

Panel A: Heterogeneous treatment effects						
Employ	Age: 21 - 35		Age: 36 - 50		Age: 51 - 64	
Estimator for θ_{w0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		0.048		-0.001		-0.029
		(0.053)		(0.037)		(0.039)
GMM-Lasso		0.042		-0.030		-0.062
		(0.086)		(0.067)		(0.075)
DRSMD-Lasso	0.095	0.047	0.016	-0.029	0.091	0.030
	(0.067)	(0.061)	(0.057)	(0.045)	(0.067)	(0.049)
Estimator for θ_{wx0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-0.080		0.046		0.073
		(0.092)		(0.083)		(0.090)
GMM-Lasso		-0.039		0.086		0.191
		(0.123)		(0.122)		(0.148)
DRSMD-Lasso	-0.147***	-0.042	-0.027	0.092	-0.074**	0.074
	(0.034)	(0.099)	(0.029)	(0.092)	(0.037)	(0.100)
Panel B: Homogeneous treatment effects						
GMM	0.012	0.024	0.017	0.009	0.001	-0.012
	(0.047)	(0.045)	(0.038)	(0.034)	(0.042)	(0.036)
GMM-Lasso	0.022	0.037	0.006	-0.033	0.019	-0.065
	(0.098)	(0.086)	(0.089)	(0.067)	(0.103)	(-0.065)
DRSMD-Lasso	0.027	0.025	0.006	0.007	0.059	0.064
	(0.06)	(0.063)	(0.052)	(0.053)	(0.059)	(0.061)
N	5962		6693		5917	

Table C.9: Heterogeneous Treatment Effects of Medicaid on Employ

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid × Above 50% Federal Poverty Line.

Average Treatment Effects						
Employ	Age: 21 - 35		Age: 36 - 50		Age: 51 - 64	
Estimator for LATE	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		0.000		0.026		0.017
		(0.052)		(0.047)		(0.054)
GMM-Lasso		0.018		0.020		0.058
		(0.105)		(0.103)		(0.127)
DRSMD-Lasso	0.008	0.022	0.000	0.024	0.044	0.076
	(0.059)	(0.066)	(0.050)	(0.065)	(0.057)	(0.073)
N	5962		6693		5917	

Table C.10: Heterogeneous Treatment Effects of Medicaid on Employ

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid × Above 50% Federal Poverty Line. The expression for LATE is $\theta_{w0} + \theta_{wx0}E(X)$.

Panel A: Heterogeneous treatment effects										
Employ	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for θ_{w0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		0.134*		-0.042		0.007		-0.063		0.022
		(0.071)		(0.067)		(0.048)		(0.040)		(0.064)
GMM-Lasso		0.143		-0.125		-0.069		-0.105		-0.063
		(0.112)		(0.118)		(0.089)		(0.077)		(0.123)
DRSMD-Lasso	0.255***	0.159**	0.014	-0.067	0.010	-0.001	0.033	-0.058	0.086	0.072
	(0.093)	(0.081)	(0.087)	(0.085)	(0.071)	(0.060)	(0.066)	(0.049)	(0.103)	(0.086)
Estimator for θ_{wx0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-0.070		0.066		-0.025		0.159		-0.046
		(0.128)		(0.120)		(0.100)		(0.104)		(0.127)
GMM-Lasso		-0.014		0.198		0.120		0.304*		0.157
		(0.179)		(0.145)		(0.150)		(0.162)		(0.193)
DRSMD-Lasso	-0.233***	-0.003	-0.092**	0.073	0.006	0.040	-0.083**	0.169	-0.018	0.012
	(0.043)	(0.145)	(0.039)	(0.135)	(0.037)	(0.113)	(0.036)	(0.113)	(0.041)	(0.153)
Panel B: Homogeneous treatment effects										
GMM	0.105	0.115	-0.012	-0.023	-0.003	0.001	-0.007	-0.035	-0.001	0.007
	(0.063)	(0.060)	(0.061)	(0.058)	(0.047)	(0.043)	(0.044)	(0.038)	(0.063)	(0.056)
GMM-Lasso	0.135	0.142	-0.013	-0.110	-0.018	-0.068	0.011	-0.115	0.021	-0.059
	(0.132)	(0.112)	(0.137)	(0.118)	(0.111)	(0.089)	(0.104)	(0.076)	(0.166)	(0.124)
DRSMD-Lasso	0.162**	0.158*	-0.032	-0.026	0.013	0.015	0.003	0.005	0.078	0.079
	(0.083)	(0.086)	(0.085)	(0.088)	(0.066)	(0.067)	(0.059)	(0.060)	(0.099)	(0.105)
N	3504		3596		3941		4599		2932	

Table C.11: Heterogeneous Treatment Effects of Medicaid on Employ

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid × Above 50% Federal Poverty Line.

Average Treatment Effects										
Employ	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for LATE	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		0.094		-0.001		-0.007		0.029		-0.009
		(0.071)		(0.071)		(0.056)		(0.060)		(0.077)
GMM-Lasso		0.135		0.000		0.000		0.071***		0.043
		(0.146)		(0.143)		(0.127)		(0.013)		(0.185)
DRSMD-Lasso	0.122	0.157*	-0.044	-0.021	0.014	0.021	-0.015	0.040	0.074	0.080
	(0.079)	(0.095)	(0.085)	(0.093)	(0.064)	(0.078)	(0.056)	(0.076)	(0.099)	(0.116)
N	3504		3596		3941		4599		2932	

Table C.12: Heterogeneous Treatment Effects of Medicaid on Employ

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid × Above 50% Federal Poverty Line. The expression for LATE is $\theta_{w0} + \theta_{wx0}E(X)$.

Panel A: Heterogeneous treatment effects										
Employ	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for θ_{w0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		0.211***		-0.072		0.056		-0.059		-0.171
		(0.073)		(0.075)		(0.057)		(0.050)		(0.106)
GMM-Lasso		0.196		-0.150		0.011		-0.154		-0.331**
		(0.133)		(0.116)		(0.076)		(0.087)		(0.156)
DRSMD-Lasso	0.183**	0.081	-0.028	-0.321**	-0.017	-0.150**	-0.011	0.566	0.013	0.001
	(0.072)	(0.176)	(0.079)	(0.138)	(0.065)	(0.061)	(0.059)	(0.529)	(0.107)	(0.106)
Estimator for θ_{wx0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-0.002*		0.001		-0.001		0.001		0.003
		(0.001)		(0.001)		(0.001)		(0.001)		(0.002)
GMM-Lasso		-0.001		0.002		0.000		0.003		0.006
		(0.001)		(0.001)		(0.001)		(0.002)		(0.004)
DRSMD-Lasso	0.000	-0.003	0.000	0.003	0.000	0.007***	0.000	-0.029	0.000	0.001
	(0.000)	(0.004)	(0.000)	(0.002)	(0.000)	(0.002)	(0.000)	(0.026)	(0.000)	(0.002)
N	3504		3596		3941		4599		2932	

Table C.13: Heterogeneous Treatment Effects of Medicaid on employment

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid× hhincome.

Average Treatment Effects										
Employ	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for LATE	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		0.093		-0.007		-0.039		0.024		0.127
		(0.070)		(0.077)		(0.067)		(0.068)		(0.141)
GMM-Lasso		0.147		-0.025		-0.013		0.095		0.202
		(0.141)		(0.139)		(0.115)		(0.130)		(0.288)
DRSMD-Lasso	0.169*	-0.116	-0.046	-0.057	0.000	0.345**	-0.023	-1.679	0.046	0.070
	(0.087)	(0.278)	(0.083)	(0.137)	(0.064)	(0.152)	(0.059)	(1.433)	(0.096)	(0.159)
N	3504		3596		3941		4599		2932	

Table C.14: Heterogeneous Treatment Effects of Medicaid on employment

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid× hhincome. The expression for LATE is $\theta_{w0} + \theta_{wx0}E(X)$.

Panel A: Heterogeneous treatment effects										
Employ	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for θ_{w0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		0.113*		-0.014		-0.007		-0.008		0.000
		(0.064)		(0.060)		(0.047)		(0.044)		(0.063)
GMM-Lasso		0.139		-0.017		-0.010		0.029		0.009
		(0.140)		(0.130)		(0.101)		(0.103)		(0.166)
DRSMD-Lasso	0.172**	0.199	-0.042	-0.097	-0.011	-0.007	-0.019	-0.019	0.033	0.033
	(0.084)	(0.129)	(0.081)	(0.181)	(0.063)	(0.063)	(0.058)	(0.058)	(0.099)	(0.099)
Estimator for θ_{wx0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		0.323		-1.006		0.199		-0.513		-1.173***
		(0.876)		(1.033)		(0.558)		(1.510)		(0.378)
GMM-Lasso		0.361		-1.490		0.160		-0.432		-0.574***
		(0.979)		(1.113)		(0.882)		(1.703)		(0.049)
DRSMD-Lasso	-0.029	-3.115	-0.018	4.228	0.333**	-0.331	0.070	0.294	-0.475	-0.574
	(0.129)	(7.497)	(0.163)	(9.724)	(0.159)	(1.291)	(0.480)	(1.229)	(0.456)	(0.492)
N	3504		3596		3941		4599		2932	

Table C.15: Heterogeneous Treatment Effects of Medicaid on employment

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid× TANF.

Average Treatment Effects										
Employ	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for LATE	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		0.126*		-0.057		-0.003		-0.012		-0.003
		(0.073)		(0.074)		(0.048)		(0.045)		(0.063)
GMM-Lasso		0.154		-0.081		-0.006		0.025		0.008
		(0.148)		(0.140)		(0.102)		(0.106)		(0.166)
DRSMD-Lasso	0.171**	0.070	-0.042	0.087	-0.003	-0.014	-0.018	-0.016	0.031	0.031
	(0.084)	(0.231)	(0.082)	(0.275)	(0.063)	(0.066)	(0.058)	(0.057)	(0.099)	(0.099)
N	3504		3596		3941		4599		2932	

Table C.16: Heterogeneous Treatment Effects of Medicaid on employment

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid× TANF. The expression for LATE is $\theta_{w0} + \theta_{wx0}E(X)$.

Panel A: Heterogeneous treatment effects										
Employ	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for θ_{w0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		0.121		-0.047		-0.046		-0.008		-0.030
		(0.170)		(0.146)		(0.103)		(0.098)		(0.163)
GMM-Lasso		0.185		-0.089		-0.017		-0.040		-0.040
		(0.267)		(0.216)		(0.138)		(0.153)		(0.263)
DRSMD-Lasso	0.282***	0.193	0.000	-0.054	0.005	-0.026	-0.002	-0.024	0.017	-0.096
	(0.107)	(0.192)	(0.099)	(0.168)	(0.088)	(0.114)	(0.069)	(0.109)	(0.125)	(0.188)
Estimator for θ_{wx0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-0.002		0.009		0.020		-0.001		0.012
		(0.070)		(0.064)		(0.048)		(0.046)		(0.068)
GMM-Lasso		-0.017		0.021		0.004		0.033		0.021
		(0.102)		(0.087)		(0.063)		(0.068)		(0.101)
DRSMD-Lasso	-0.047**	-0.022	-0.018	0.005	-0.005	0.007	-0.008	0.000	0.006	0.043
	(0.022)	(0.074)	(0.020)	(0.070)	(0.020)	(0.051)	(0.017)	(0.050)	(0.023)	(0.075)
N	3504		3596		3941		4599		2932	

Table C.17: Heterogeneous Treatment Effects of Medicaid on employment

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid× smoke.

Average Treatment Effects										
Employ	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for LATE	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		0.118*		-0.025		-0.003		-0.009		-0.001
		(0.065)		(0.063)		(0.048)		(0.045)		(0.064)
GMM-Lasso		0.144		-0.040		-0.007		0.031		0.011
		(0.142)		(0.131)		(0.105)		(0.104)		(0.168)
DRSMD-Lasso	0.171**	0.142*	-0.043	-0.044	-0.007	-0.010	-0.020	-0.024	0.032	0.006
	(0.084)	(0.083)	(0.081)	(0.080)	(0.062)	(0.064)	(0.058)	(0.058)	(0.099)	(0.096)
N	3504		3596		3941		4599		2932	

Table C.18: Heterogeneous Treatment Effects of Medicaid on employment

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid× smoke. The expression for LATE is $\theta_{w0} + \theta_{wx0}E(X)$.

Panel A: Heterogeneous treatment effects						
Out of Pocket Cost	Age: 21 - 35		Age: 36 - 50		Age: 51 - 64	
Estimator for θ_{w0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-249.086***		-94.061*		-171.578***
		(73.769)		(50.838)		(53.230)
GMM-Lasso		-465.791***		-236.344***		-320.039***
		(110.733)		(95.697)		(96.416)
DRSMD-Lasso	-250.542**	-309.857***	-75.370	-67.352	1.626	-162.346**
	(102.629)	(90.142)	(87.700)	(65.262)	(114.941)	(67.412)
Estimator for θ_{wx0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		127.498		-109.671		193.479
		(141.823)		(128.054)		(144.174)
GMM-Lasso		295.686		-63.730		254.041
		(187.704)		(188.740)		(261.915)
DRSMD-Lasso	29.129	163.879	-46.060	-60.297	-112.079**	263.280
	(52.022)	(149.725)	(35.699)	(142.875)	(55.790)	(170.840)
Panel B: Homogeneous treatment effects						
GMM	-191.55***	-210.785***	-136.74**	-118.182**	-91.278	-127.007**
	(68.965)	(63.796)	(57.490)	(48.792)	(64.973)	(52.175)
GMM-Lasso	-323.057**	-431.615***	-263.160**	-234.630**	-211.509	-323.868***
	(141.975)	(112.434)	(133.556)	(94.993)	(170.007)	(95.266)
DRSMD-Lasso	-237.074***	-222.744***	-92.681	-90.985	-46.131	-41.349
	(91.39)	(95.524)	(80.402)	(81.631)	(100.07)	(103.067)
N	5962		6693		5917	

Table C.19: Heterogeneous Treatment Effects of Medicaid on Out of Pocket Cost

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid × Above 50% Federal Poverty Line.

Average Treatment Effects						
Out of Pocket Cost	Age: 21 - 35		Age: 36 - 50		Age: 51 - 64	
Estimator for LATE	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-173.233**		-157.835**		-50.003
		(79.133)		(74.153)		(88.032)
GMM-Lasso		-289.877*		-273.403*		-160.41
		(156.242)		(156.520)		(217.11)
DRSMD-Lasso	-233.213***	-212.36**	-102.154	-102.415	-68.800	3.088
	(89.090)	(100.468)	(77.14)	(100.419)	(94.184)	(126.818)
N	5962		6693		5917	

Table C.20: Heterogeneous Treatment Effects of Medicaid on Out of Pocket Cost

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid × Above 50% Federal Poverty Line. The expression for LATE is $\theta_{w0} + \theta_{wx0}E(X)$.

Panel A: Heterogeneous treatment effects										
Out of Pocket Cost	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for θ_{w0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-238.166**		-294.692***		-32.148		-149.770***		-153.110*
		(94.947)		(92.333)		(69.362)		(53.074)		(87.341)
GMM-Lasso		-545.537***		-569.684***		-276.342**		-321.387***		-465.763**
		(148.534)		(141.359)		(131.711)		(91.886)		(189.849)
DRSMD-Lasso	-373.155***	-250.372**	-153.656	-376.780***	-55.406	18.780	-162.227	-185.345***	262.158	-49.106
	(136.072)	(119.078)	(142.101)	(126.053)	(99.185)	(84.691)	(101.883)	(68.418)	(197.082)	(114.506)
Estimator for θ_{wx0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-101.939		384.290**		-269.238*		155.502		173.800
		(193.546)		(181.372)		(156.752)		(165.046)		(211.235)
GMM-Lasso		226.006		558.392**		-77.050		307.058		477.385
		(245.546)		(243.326)		(209.610)		(276.655)		(370.165)
DRSMD-Lasso	104.666*	-179.667	46.142	496.881**	-72.394*	-242.417	6.897	72.042	-111.454*	502.058*
	(62.096)	(214.131)	(53.675)	(214.081)	(43.318)	(167.287)	(45.935)	(181.569)	(57.602)	(295.269)
Panel B: Homogeneous treatment effects										
GMM	-281.393***	-265.709***	-118.835	-183.904**	-140.436*	-100.078	-94.476	-122.298**	-68.536	-98.839
	(89.881)	(82.886)	(90.944)	(82.150)	(72.634)	(64.576)	(67.326)	(53.027)	(100.771)	(82.629)
GMM-Lasso	-447.852**	-532.224***	-255.052	-527.148***	-308.891*	-276.433**	-204.110	-330.655***	-210.799	-453.139**
	(176.244)	(142.177)	(206.913)	(151.105)	(159.359)	(131.643)	(163.024)	(89.334)	(302.765)	(192.478)
DRSMD-Lasso	-331.473***	-330.392***	-130.819	-98.119	-83.096	-81.152	-159.723*	-158.614*	208.194	232.779
	(119.712)	(123.117)	(138.478)	(145.535)	(91.477)	(93.811)	(92.494)	(93.832)	(186.861)	(198.671)
N	3504		3596		3941		4599		2932	

Table C.21: Heterogeneous Treatment Effects of Medicaid on Out of Pocket Cost

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid × Above 50% Federal Poverty Line.

Average Treatment Effects										
Out of Pocket Cost	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for LATE	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-296.380***		-52.748		-187.091**		-59.762		-36.809
		(104.514)		(108.950)		(89.257)		(95.636)		(128.929)
GMM-Lasso		-416.474**		-218.127		-320.684*		-143.655		-146.312
		(197.893)		(220.931)		(179.511)		(212.015)		(347.923)
DRSMD-Lasso	-313.384***	-352.973***	-124.606	-63.950	-97.068	-120.728	-158.234*	-143.646	187.576	286.854
	(113.461)	(137.349)	(138.286)	(154.905)	(88.592)	(110.318)	(87.948)	(122.867)	(183.656)	(227.628)
N	3504		3596		3941		4599		2932	

Table C.22: Heterogeneous Treatment Effects of Medicaid on Out of Pocket Cost

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid × Above 50% Federal Poverty Line. The expression for LATE is $\theta_{w0} + \theta_{wx0}E(X)$.

Panel A: Heterogeneous treatment effects										
Out of Pocket Cost	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for θ_{w0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-157.173* (95.566)		-338.762*** (115.896)		74.768 (99.214)		-197.276** (76.924)		-120.243 (198.875)
GMM-Lasso		-438.811*** (163.018)		-605.793*** (177.487)		-260.345** (123.764)		-526.900*** (122.174)		-483.119 (305.685)
DRSMD-Lasso	-309.745*** (102.007)	-145.185 (260.982)	-187.617 (125.011)	-237.79 (204.272)	-81.096 (88.358)	58.587 (87.128)	-182.139** (90.764)	-949.979 (663.448)	213.045 (195.506)	-201.021 (123.915)
Estimator for θ_{wx0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-2.089 (1.294)		3.889* (2.183)		-4.594* (2.486)		2.234 (1.992)		0.994 (4.800)
GMM-Lasso		-0.208 (1.349)		5.190* (2.858)		0.064 (2.175)		6.601** (3.138)		8.209 (8.435)
DRSMD-Lasso	-0.268 (0.543)	-6.179 (5.497)	0.737* (0.428)	-1.490 (3.529)	-0.374 (0.582)	-2.399 (2.276)	0.140 (0.377)	37.031 (31.714)	-1.034** (0.496)	5.204** (2.432)
N	3504		3596		3941		4599		2932	

Table C.23: Heterogeneous Treatment Effects of Medicaid on Out of Pocket Cost

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid× hhincome.

Average Treatment Effects										
Out of Pocket Cost	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for LATE	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-311.890**		-28.504		-260.520**		-26.410		-31.122
		(100.413)		(123.187)		(121.165)		(113.466)		(260.191)
GMM-Lasso		-454.203**		-191.738		-255.659		-22.092		252.781
		(186.231)		(219.071)		(166.964)		(218.677)		(558.109)
DRSMD-Lasso	-329.612***	-602.840**	-128.813	-356.685*	-108.424	-116.457	-171.432*	1881.963	120.352	265.458
	(126.950)	(279.909)	(132.003)	(194.470)	(92.768)	(148.272)	(90.212)	(1774.876)	(172.107)	(215.602)
N	3504		3596		3941		4599		2932	

Table C.24: Heterogeneous Treatment Effects of Medicaid on Out of Pocket Cost

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid× hhincome. The expression for LATE is $\theta_{w0} + \theta_{wx0}E(X)$.

Panel A: Heterogeneous treatment effects										
Out of Pocket Cost	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for θ_{w0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-274.917***		-127.619		-144.510		-103.605		-69.921
		(91.821)		(88.876)		(73.094)		(66.604)		(100.480)
GMM-Lasso		-408.080**		-217.524		-248.498*		-158.129		-16.772
		(187.386)		(194.537)		(146.186)		(160.274)		(302.841)
DRSMD-Lasso	-328.977***	-361.671**	-144.073	-142.501	-92.484	-115.336	-174.760*	-171.099*	158.324	158.070
	(122.904)	(145.690)	(129.297)	(169.190)	(88.281)	(89.513)	(89.460)	(90.094)	(180.997)	(181.520)
Estimator for θ_{wx0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-360.531		438.043		341.411		3124.245		160.364
		(595.971)		(1474.901)		(692.758)		(3527.129)		(681.382)
GMM-Lasso		-2695.988*		-1430.592		-751.087		3216.942		620.170
		(1567.443)		(2019.704)		(1111.768)		(4538.626)		(383.799)
DRSMD-Lasso	142.752	2647.014	67.449	3.327	-429.745	1630.753	-178.469	-1515.266	741.691	1129.482
	(97.306)	(5599.135)	(196.456)	(6734.098)	(364.850)	(1734.534)	(358.992)	(2158.009)	(699.283)	(1056.448)
Panel B: Homogeneous treatment effects										
GMM	-280.010***	-276.059***	-121.970	-126.700	-140.874*	-143.04**	-94.666	-101.783	-69.635	-69.778
	(91.172)	(91.62)	(90.067)	(88.798)	(72.829)	(72.905)	(67.076)	(66.588)	(100.400)	(100.438)
GMM-Lasso	-454.731**	-389.029**	-239.756	-211.541	-256.614*	-253.578*	-156.003	-176.968	-15.345	-22.411
	(187.884)	(186.352)	(198.492)	(194.269)	(145.706)	(145.766)	(160.649)	(158.521)	(303.056)	(302.061)
DRSMD-Lasso	-326.593***	-332.779***	-143.005	-142.455	-98.117	-101.027	-175.448**	-175.753**	160.566	162.147
	(122.625)	(122.35)	(129.513)	(128.689)	(88.763)	(88.614)	(89.326)	(89.209)	(181.463)	(181.422)
N	3504		3596		3941		4599		2932	

Table C.25: Heterogeneous Treatment Effects of Medicaid on Out of Pocket Cost

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid× TANF.

Average Treatment Effects										
Out of Pocket Cost	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for LATE	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-289.837***		-108.616		-136.800*		-75.073		-69.429
		(92.479)		(107.224)		(73.492)		(73.867)		(100.347)
GMM-Lasso		-519.644***		-279.585		-265.460*		-128.750		-14.868
		(202.286)		(217.774)		(146.215)		(171.223)		(303.081)
DRSMD-Lasso	-323.070***	-252.134	-141.147	-142.357	-102.189	-78.509	-176.390**	-184.937**	160.600	161.537
	(122.184)	(197.461)	(130.029)	(221.220)	(89.195)	(93.360)	(89.177)	(89.035)	(181.463)	(181.433)
N	3504		3596		3941		4599		2932	

Table C.26: Heterogeneous Treatment Effects of Medicaid on Out of Pocket Cost

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid× TANF. The expression for LATE is $\theta_{w0} + \theta_{wx0}E(X)$.

Panel A: Heterogeneous treatment effects										
Out of Pocket Cost	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for θ_{w0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-534.801*		-285.780		-263.638*		-241.214		-254.837
		(278.120)		(225.302)		(144.724)		(146.954)		(241.115)
GMM-Lasso		-555.430		-374.526		-440.806***		-566.475**		-578.634
		(361.855)		(341.910)		(167.480)		(230.489)		(427.954)
DRSMD-Lasso	-425.270***	-685.989**	-165.238	-521.033**	-62.497	-236.941	-126.876	-338.461**	326.003	-214.314
	(159.063)	(299.677)	(162.906)	(264.138)	(125.410)	(171.558)	(105.255)	(164.009)	(218.319)	(299.180)
Estimator for θ_{wx0}	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		109.750		74.501		60.194		71.647		80.074
		(108.556)		(99.719)		(73.097)		(70.949)		(105.337)
GMM-Lasso		43.733		57.688		93.144		197.556*		247.086
		(138.335)		(137.141)		(93.057)		(109.936)		(176.298)
DRSMD-Lasso	42.055	151.033	9.669	151.900	-17.395	74.186	-23.615	70.550	-71.546**	160.300
	(29.563)	(111.229)	(28.358)	(110.100)	(26.269)	(77.875)	(20.306)	(76.311)	(31.867)	(119.734)
N	3504		3596		3941		4599		2932	

Table C.27: Heterogeneous Treatment Effects of Medicaid on Out of Pocket Cost

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid× smoke.

Average Treatment Effects										
Out of Pocket Cost	Age: 21 - 29		Age: 30 - 38		Age: 39 - 47		Age: 48 - 56		Age: 57 - 64	
Estimator for LATE	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)	Z_1	(Z_1, Z_1X_1)
GMM		-276.588*** (90.822)		-112.372 (92.812)		-133.077* (76.291)		-87.451 (68.967)		-62.898 (102.980)
GMM-Lasso		-452.537** (188.316)		-240.252 (198.422)		-238.778 (155.644)		-142.495 (165.046)		13.630 (313.414)
DRSMD-Lasso	-326.325*** (122.502)	-330.648*** (120.503)	-142.732 (129.249)	-167.472 (127.050)	-100.227 (87.164)	-76.031 (89.525)	-177.557** (89.010)	-187.052** (91.782)	154.507 (180.620)	169.926 (168.551)
N	3504		3596		3941		4599		2932	

Table C.28: Heterogeneous Treatment Effects of Medicaid on Out of Pocket Cost

Note: *** Significant at 1%, ** at 5%, * at 10%. Each row shows the estimates and robust standard errors for the same type of estimator. In the columns, we present the instruments these estimators used and the age groups. The interaction term is Medicaid× smoke. The expression for LATE is $\theta_{w0} + \theta_{wx0}E(X)$.