University of Mannheim
Department of Economics

Doctoral Thesis

Essays in Nonparametric Instrumental Variable Regression

Petyo Bozhidarov Bonev

Supervisors: Professor Gerard J. van den Berg, Professor Enno Mammen

Mannheim, September the 8th, 2014

Referenten: Professor Dr. Gerard van den Berg, Professor Dr. Paul Gans, Professor Dr. Enno Mammen, Professor Dr. Volker Nocke

Abteilungssprecher: Professor Dr. Eckhard Janeba

Datum der Prüfung: 24.09.2014

Acknowledgements

I am greatly indebted to Gerard van den Berg and Enno Mammen for their guidance and supervision, patience and critical assessment of my work. I learned a lot about econometrics and its applications from them. I also benefited greatly from numerous discussions with Christoph Breunig and Martin Wahl. Further, I am thankful to Markus Frölich, Stephen Kastroyano, Andreas Landmann and all other colleagues from the department of economics at the University of Mannheim who never closed the door when I passed by to discuss my ideas. I also thank to Sylvie Blasco, Bettina Drepper, Bo E. Honoré, Aureo de Paula, Gautam Tripathi and participants at the ESEM, an IZA conference on labor market policy evaluation at Harvard, conferences on survival analysis and on the evaluation of political reforms at Mannheim, a workshop at ZEW, and the joint econometrics and statistics workshop at the LSE, for their useful comments. The access to the data used in the first chapter has been made possible thanks to INSEE-CREST, DARES at the French Ministry of Labor and Pôle Emploi, but especially to Bruno Crépon, Thomas le Barbanchon, Francis Kramarz, and Philippe Scherrer, who made the project possible with their extraordinary help and hospitality. I am also thankful to Paul Glewwe and Boris Kramarz for giving me access to their data sets for the second chapter of my thesis. And last but not least, I would like to thank to my wife Lena Boneva, who not only had enough patience and understanding with me, but contributed actively by giving useful comments at each stage of my thesis.

General introduction

Endogeneity hampers the econometric evaluation of economic causal relationships. Broadly speaking, an economic model is affected by endogeneity when unobserved factors influencing the dependent variable of interest are also related to (observed) independent variables. In such cases, the causal effect of the observables cannot be separated from the causal effect of the unobservables even if there are infinitely many data observations. Endogeneity is the result of the complexity of human behavior and of the impossibility to observe all relevant information. It is inherent in all economic disciplines. In labor economics for example, the level of intelligence of an individual is generally unobservable and is related to the level of education and other predictors of economic success. In the economics of education, school resources, such as the quality of teachers or class sizes, might be related to the economic background of students or the level of parents' support. The latter factor of school achievement is often unobserved to the econometrician. In health studies, researchers are often concerned with the dependence between genetic factors and behavior when they try to predict the individual economic wellbeing or the accumulation of human capital.

These are only three out of many examples that demonstrate that the importance of methods that can solve the problem of endogeneity cannot be overstated. When experiments are not possible (due to political, financial or ethical reasons), instrumental variable methods are one major potential solution to it. Instrumental variables (or simply instruments) are variables that i) are observable, ii) are related to the observed independent variables and iii) influence the dependent variable only through the independent ones. Variation in the instruments can be used to "extract" the exogenous variation of the observed covariates, which is then used to estimate the causal relationship of interest.

In my thesis, I contribute to the literature on instrumental variable(IV) methods in several ways. First, I develop a new nonparametric IV method for treatment evaluation in the context of duration models. Second, I provide a characterization of a broad class of nonparametric penalized minimum distance IV estimators as projections and derive results about their asymptotic properties. These contributions are the subjects of the two different chapters of my thesis. I now give a brief outline of the structure and particular content of my work.

Chapter 1 of my thesis is based on the paper "Nonparametric instrumental variable methods for dynamic treatment evaluation", a joint project with Gerard Van den Berg and Enno Mammen. The main object of interest is the distribution (or some functions of it) of a duration variable. In a policy treatment evaluation framework, our identification and estimation methods allow for two types of endogeneity. The first type arises from the decision of agents based on unobservables to take or refuse an assigned treatment. The second type arises over time due to selective exits of agents out of the population of interest. Both types of endogeneity are inherent to economic policy evaluation but there is no method thus far that tackles them both at the same time in a nonparametric way. Existing methods either ignore one of the types of endogeneity or impose restrictive parametric structure. Our instrumental approach can deal with both types of endogeneity in a completely nonparametric way. It relies on random inflow into the population of interest. Furthermore, we do not assume separability or independence of observed and unobserved covariates. In addition, our methods can deal with censoring of the duration variable. We provide estimation procedures and derive their asymptotics. In addition to the identification and estimation results, we also demonstrate how to use our framework for the analysis of endogeneity.

The second chapter of my thesis is based on my paper "The effects of class size on school performance: a nonparametric study with new shape-constrained instrumental variable methods". The main econometric focus of this chapter is the nonparametric shape analysis of the mean regression function under endogeneity. Imposing shape constraints in estimation has two main advantages. First, when a certain shape is predicted by economic theory, imposing shape constraints on the estimates may be necessary for the interpretation of the data in a policy evaluation context. Second, shape-constrained nonparametric estimators are still much more flexible than parametric counterparts. An important question in the econometric analysis in this context is what are the asymptotic properties of constrained nonparametric estimators. I provide three main theoretical results. First, monotonically constrained and unconstrained Tikhonov estimators are asymptotically equivalent when the regression function is an inner point of the constrained set. In this case, imposing a constraint does not lead asymptotically to a change in the behavior of the estimator. Second, a broad class of penalized minimum distance estimators can be shown to be the projection of the unconstrained counterparts on the constrained set. An important consequence is that in some weak norm the constrained estimators, provided they exist, converge at least as quickly to the model solution as the unconstrained estimators. For a certain subclass of those estimators it can be further shown, that they are two-step projection estimators in the following sense. In a first step, project in some general vector space the data on the set of all potential regression functions to obtain the unconstrained estimator. In a second step, project this projection on the constrained set to obtain the unconstrained estimator. The third result is a demonstration of an application of the projection property. Consistency of constrained estimators is shown to be related to properties of the model solution and the conditional expectation operator. In addition, I suggest an empirical procedure for testing for monotonicity.

Both chapters of my thesis contain simulations of the proposed methods. They differ substantially in their purpose. In chapter 1, we show through a simulation study that in a particular policy context, violations of the assumption of independent censoring of the duration variable do not influence the performance of the estimator. This finding is of a great importance for applied research as the independent censoring assumption is not testable. In section 2, I derive through a simulation study the optimal choices of the estimation parameters in a sieves estimation approach and monotonicity constraint. These choices depend on the sample size, the degree of endogeneity, the strength of the instrument and the functional form of the regression function. The second part of the simulation study also shows that the proposed ad hoc testing procedure works very well in finite samples.

Both the methodological and simulation parts of my thesis have clear application motivations. Their usage is demonstrated in two extensive empirical investigations in two important economic disciplines: labor and education. In chapter 1, we evaluate a labor policy reform that i) introduces active labor market policy measures for unemployed, such as training and counselling, and ii) abolishes digression in unemployment insurance payments. The two measures induce incentives with opposite direction and it is not clear what the overall impact on the duration of unemployment would be. We use our IV approach for the evaluation and find that the reform had a positive effect on the unemployment dynamics. In chapter 2, I find that the effect of class size on test scores in two different data sets is non-monotone. Building on that novel finding, I suggest a simple educational production function that can generate non-monotone causal effects.

The last result completes the chain econometric theory - simulation - empirical investigation - economic theory. It reflects my understanding of an integrated scientific process and demonstrates that I have developed a broad set of skills during my PhD time.

Contents

Chapter 1

Nonparametric instrumental variable methods for dynamic treatment evaluation

1.1 Introduction

Identification of dynamic treatment effects is hampered by three major problems. First, suppose the treatment is randomized at $t = 0$, the point in time of inflow of individuals in some state of interest. If unobservable factors of the outcome interact with the treatment status, then, at some later point in time $t > 0$, the distributions of the unobserved characteristics among survivors will differ across different treatment arms, Meyer (1996), Ham and LaLonde (1996) and Eberwein, Ham, and LaLonde (1997). Additionally, if individuals can choose a treatment status different from the one that has been assigned to them and if their decision is related to unobserved characteristics, then estimation results will suffer from the standard selection bias. We refer to these endogeneity causes as dynamic and static endogeneity. Lastly, duration variables are often subject to censoring, a problem which is difficult to tackle with standard regression methods. In this paper, we develop an instrumental variable (IV) approach for identification and estimation of dynamic treatment effects on the conditional survival function and the hazard of a duration variable. Our method solves the dynamic and static endogeneity problems and allows for censoring. We do not adopt parametric or semiparametric structure. We also do not impose independence or separability of observed and unobserved characteristics.

We embed our IV approach in a dynamic regression discontinuity setting. A single comprehensive treatment is assigned at a specific calendar point in time to all individuals in some state of interest. A typical example is a labor market reform which changes the structure of unemployment benefits. Cohorts of individuals receive the treatment at the same point in time but at different elapsed durations of their spells. Due to dynamic selection the distribution of unobserved characteristics at the moment of treatment will differ across cohorts. Additionally, we allow for noncompliance. We achieve identification by using the duration between inflow and treatment of the different cohorts as an instrument for the endogenous dynamic treatment status. The identifying assumptions are that different cohorts have equal distributions of the frailty at the moment of inflow conditional on observables

and that individuals do not anticipate the point in time of treatment or do not act upon this information.

Additionally, our identification strategy can be applied to a setup in which individual spells have the same starting point in time but the agents receive the treatment at different (random) points in time. The latter setting is common to the Swedish practice of Active Labor Market Policies (ALMP), see Sianesi (2004).

By dealing with both dynamic and static selection, our paper provides the link between the standard (static) LATE literature and the literature on dynamic treatment evaluation. On the one hand, our main result can be interpreted as a dynamic generalization of the one-sided noncompliance identification result by Bloom (1984). On the other hand, our strategy generalizes the method of Van den Berg, Bozio, and Dias (2014) by allowing for static selection.

We suggest estimation procedures and derive their asymptotic properties. Our estimators are dynamic versions of the Wald estimator.

We use our method to evaluate the French labor market reform PARE from 2001. On July 1 2001 the digression of the unemployment benefits over time was abolished and a package of active labor market policy (ALPM) measures was introduced. The estimated treatment effect of this reform on the conditional survival function is positive and increases over time. In an exhaustive study, we defend the plausibility of our assumptions. We address the non-testable random censoring assumption in a simulation study. Imposing of random censoring is necessary due to a nonidentification result by Tsiatis (1975). Our simulation results indicate that the estimator is robust to violations of the non-testable assumption. The reason is that violations which are likely to occur in the PARE setting have opposite directions and offset each other's impact on the estimates. This is a novel result.

Finally, we provide a novel framework for analysis of endogeneity. The main purposes are 1) to assess whether noncompliance is endogenous and 2) to measure the bias that would be induced if the endogeneity is ignored. Understanding the selection process is important in numerous economic applications. First, better knowledge of the reasons for the non-take up of a policy reform help improve the policy design and increase its efficiency. Second, evaluating pilot projects with noncompliance can be used to derive bounds for the effect of a comprehensive policy reform (with perfect compliance). And third, better understanding of endogeneity can be used to model selection explicitly in more complex models. Our methods are based on a comparison of untreated noncompliers with a whole nontreated cohort at the same elapsed duration. We evaluate the non-take up of the PARE reform. Our results indicate that selection is endogenous and that one major reason for noncompliance is the expectation of a quick exit. These findings are in line with previous empirical and theoretical studies, see e.g. Blasco (2009).

The remainder of this paper is structured as follows. In section 2, we discuss the related literature. We present our IV approach in section 3. In section 4, we apply our IV method to the French labor market reform PARE. Section 6 concludes. All proofs are left for the appendix.

1.2 Literature overview

The related literature can be divided into a theoretical and empirical strands. Our identification approach is related to the theoretical literature on regression discontinuity design. Some recent developments in this field are those by Hahn, Todd, and van der Klaauw (2001), Porter (2003), Frölich (2007) and Van den Berg, Bozio, and Dias (2014). The first three approaches do not incorporate censoring. The paper of Van den Berg, Bozio, and Dias (2014) deals with censoring and dynamic selection. We generalize their model by allowing for noncompliance. Our IV approach is related to the IV methods with a binary treatment such as those in Imbens and Angrist (1994) and Imbens and Rubin (1997). These papers pose the analysis in a static framework and do not consider censoring and dynamic endogeneity. IV methods for duration data are considered for example in Robins and Tsiatis (1991), Chesher (2002), Bijwaard and Ridder (2005), Bijwaard (2008) and Abbring and van den Berg (2005). Typically, these studies adopt a semiparametric or a parametric structure. In their numerous settings, Abbring and van den Berg (2005) either preclude dynamic selection by looking at the unconditional survival function or adopt a semiparametric structure. Next, our paper is related to the literature on dynamic matching estimators, see e.g. Sianesi (2004), Fredriksson and Johansson (2008) and Crépon, Ferracci, Jolivet, and Van den Berg (2009), and to the literature on dynamic discrete choices, see e.g. Heckman and Navarro (2007). These papers assume full compliance. We discuss in detail their assumptions and results in section 1.3.2.

On the empirical side, we contribute to the literature on the influence of the structure of unemployment insurance benefits on the unemployment duration, see for example Lalive (2008), Lalive, van Ours, and Zweimüller (2006) and Katz and Meyer (1990). Commonly, the unemployment insurance expires after some predetermined period of time. This has driven the literature to consider the impact of maximal length of the period of payments and the amount of the (flat) entitlement on the unemployment duration hazard. Comparing flat with digressive benefits, we contribute to this literature by giving insights on the influence of the *interim* structure of the unemployment insurance payments on the employment dynamics. A related question is studied in Prieto (2000) and Dormont, Fougère, and Prieto (2001), but their econometric approaches involve the (semi-)parametric specifications of the Proportional Hazards model and the Mixed Proportional Hazards model, respectively, which are hard to justify with economic theory, Van den Berg (2001). Our methods avoid such restrictive assumptions and rely solely on the timing of the treatment. Our paper contributes also to the empirical literature on the effects of ALMP on the probability to find a job. The measures introduced by the PARE reform include training, subsidized jobs, skill assessment and job search assistance. Some studies considering training are Gritz (1997), Richardson and den Berg (2001), Crépon, Ferracci, and Fougère (2007) and Crépon, Ferracci, Jolivet, and Van den Berg (2009). Studies on the effectiveness of counselling can be found in Gorter and Kalb (1996), Blundell, Dias, Meghir, and Reenen (2004), as well as in Crépon, Dejemeppe, and Gurgand (2005), Van den Berg and Van der Klaauw (2010) and Van den Berg, Kjærsgaard, and Rosholm (2012). Studies considering subsidized jobs are Gerfin and Lechner (2002) and Blundell, Dias, Meghir, and Reenen (2004). For a general overview see Bonnal, Fougère, and Sérandon (1997), Heckman, LaLonde, and Smith (1999) and Kluve (2010). A common feature of most of these studies is the assumed parametric or semiparametric functional form of the hazard. Lastly, another related branch of the literature focuses on the threat effects of ALMP. Some recent papers are those of Black, Smith, Berger, and Noel (2003), Lalive, Zweimüller,

and van Ours (2005), as well as Rosholm and Svarer (2008), Crépon, Ferracci, Jolivet, and Van den Berg (2010) and Bergemann, Caliendo, van den Berg, and Zimmermann (2011).

1.3 Identification and estimation of dynamic treatment effects

1.3.1 Notation and a framework for dynamic treatment evaluation

Assume that all agents in some state of interest O are assigned to receive a treatment at a specific calender point in time $r > 0$. We are interested in the causal effect of this treatment on the distribution of the duration of stay in O. We embed our analysis in a framework with dynamic potential outcomes. We assume that potential outcomes of the individual i depend on pretreatment characteristics X_i and V_i, of which the q-dimensional X_i is observed, $q \geq 1$, and the one-dimensional V_i not. Let the random variable Z_i denote the time from inflow to the assigned point in time of treatment and S_i the elapsed duration in O at which individual i actually receives the treatment. S_i is a choice variable whereas Z_i is exogenous. For each $X = x, V = v, Z = z, S = s$, denote with $T_i(s, z, x, v)$ the potential duration of stay in O of individual i if he or she had characteristics (x, v) and received (z, v) as values for (Z, S). We allow $T_i(s, z, x, v)$ to be a random variable. This assumption reflects some intrinsic uncertainty in the transition, not necessarily observed and/or controlled by the agent, see Lancaster (1990) for a discussion. Throughout the paper, we assume that Z is an exclusion restriction in the sense that $T_i(s, z, x, v) = T_i(s, x, v)$. For notational simplicity, we will suppress the dependence on X and V as well as the individual index i.

This setup corresponds to a labor market program implementation, in which a policy reform is administered at a fixed point in time. Our methods however, as shown in the discussion below, can be extended to a setup with ongoing programs, in which the treatment is assigned at random points in time to different individuals. The latter setting is common to the Swedish practice of Active Labor Market Policies (ALMP), see Sianesi (2004). In a labor market context, X might be education, gender, number of siblings, age and experience at inflow, whereas V might be the ability of an unemployed or his or her motivation. In a medical study, X might be some observed health marker, whereas V might be some genetic unobserved component. X and V obtain values in Ω_X and Ω_V.

We enrich this dynamic framework by allowing the agents to opt out of the assigned treatment. We refer to this opting out as static selection. To fix ideas, for each $z \in \mathbb{R}_+$ and each $(x, v) \in \Omega_X \times \Omega_V$, let the random variable $S(z, x, v)$ denote the potential compliance status of an individual with observed and unobserved characteristics x and v, respectively, given that the treatment z is assigned to that individual. For notational simplicity, we write $S(z)$. $S(z)$ can be interpreted as the potential elapsed duration in O at which an agent would like to be treated, if he or she was assigned to be treated at elapsed duration z. To make the model tractable, an agent is only allowed to accept or reject an assigned treatment, and the treatment is only offered once (see assumption A1 in the following subsection, as well as the corresponding discussion). Thus, for each $z \in \mathbb{R}_+$, $S(z)$ may take only the values z (the case of compliance) and ∞ (the case of noncompliance).[1] Agents are allowed to have an

[1] Alternatively, we might restrict the maximal potential duration of the state of interest to be equal to some positive real number $\bar{S}$. In that case, noncompliers receive $S(z) = \bar{S}$. We do not differentiate between these two

arbitrary time structure of their compliance preferences. A cancer suffering patient might be reluctant to accept a new therapy at an early stage of the disease, but his or her preference might change at an advanced stage of the disease. Similarly, an unemployed person might refuse a training early in the unemployment spell and be willing to attend it later on. To account for the possibility of changing preferences, we refer to individuals who would be willing to receive a treatment at some elapsed duration z, given that they were asked to do so, as z-compliers. This notion generalizes the static compliance definition. This very general framework allows us also to incorporate individual expectations about the own potential outcome at different points in time.

Allowing for static selection is common in the standard literature on (static) treatment evaluation, see Heckman and Vytlacil (2007). In a labor market program, unemployed individuals might decide not to accept an offer for a training or a counselling service. An often quoted example is the Job Training Partnership Act (JTPA) program, see Bloom, Orr, Bell, Cave, Doolittle, Lin, and Bos (1997). In a medical study, patients assigned to drop out from a therapy might be able to participate in a substitute program. Selection into or out of a certain treatment status creates a potential endogeneity problem, which has given rise to the development of the Local Average Treatment Effect (LATE) literature, see Imbens and Angrist (1994). Typically, the randomized treatment assignment is used as an instrument for the endogenous actual treatment status.[2]

While the standard LATE literature poses the evaluation problem as a static problem and the time dimension is ignored, there is a branch of the econometric literature that focuses on dynamic selection and precludes the possibility of static selection, see for example Eberwein, Ham, and LaLonde (1997),Abbring and van den Berg (2003), Heckman and Navarro (2007) for different methods of accounting for dynamic selection, as well as Abbring and Heckman (2007) for an overview of dynamic treatment evaluation methods. Dynamic selection may arise even when the experiment has been perfectly randomized at some initial point in time $t = 0$ of the state of interest. If the unobserved heterogeneity interacts with the treatment status, then its distribution at a later point in time $t > 0$ might differ between the different treatment arms due to differences in the dynamics of transitions, see also Abbring and van den Berg (2005). We develop a framework that deals with both static and dynamic types of selection. Thus, we provide the link between the two branches of literature.

Let T be the actual duration of the spell. T might be right censored by a random variable C. Define $\widetilde{T} := \min\{T, C\}$ and the censoring indicator $\delta := 1\{\widetilde{T} = T\}$. We observe $(\widetilde{T}, \delta)$ and not directly (T, C). We assume access to an i.i.d. sample

$$(\widetilde{T}_1, S_1, Z_1, X_1, \delta_1), \ldots, (\widetilde{T}_n, S_n, Z_n, X_n, \delta_n),$$

where S_i is missing if $S_i > \widetilde{T}_i$.

Remark
Unless explicitly otherwise stated, we will denote with t, s, z elapsed durations in O (and not calender time). Thus, for example, 0 refers to the point in time of inflow of an agent into O. Furthermore, we do not need a binary process $D_i(t)$ that denotes the treatment status of an agent i at time t. Before the calender point in time r, nobody is treated. After r, all compliers

cases and write ∞.

[2]In line with the biometry literature, this instrument is also called Intention-to-Treat (ITT)

are treated, that is, all individuals whose value of S is equal to the corresponding value of Z. Therefore, the treatment status can be deduced from S, Z and the calender time.

The treatment effect of interest is

$$P(T(s) \in [t, t+a) \mid T(s) \geq t', X, V) - P(T(s') \in [t, t+a) \mid T(s') \geq t', X, V), \tag{1.3.1}$$

that is, the additive effect of replacing the treatment s' with the treatment s on the probability to exit the state of interest between t and $t+a$ conditionally on surviving up to t'. The case $s' = \infty$ induces a comparison between those treated at s and those never treated. Another special case is the limit case $a \to 0$, $t' = t$. Denote with $\theta_{T(s)}(t \mid X, V)$ the hazard of $T(s)$ at t for an individual with characteristics X and V. Then the individual additive treatment on the hazard at t is defined as

$$\theta_{T(s)}(t \mid X, V) - \theta_{T(s')}(t \mid X, V). \tag{1.3.2}$$

It reflects the additive change in the exit rate induced by a change of the treatment from s' to s. Additive effects on the distribution of the potential outcome are common in the literature, see for example Fredriksson and Johansson (2008) and Crépon, Ferracci, Jolivet, and Van den Berg (2009) for an effect on the unconditional survival function, Abbring and van den Berg (2005) for an effect on the conditional survival function, Van den Berg, Bozio, and Dias (2014) for an effect on the hazard. One appealing feature of additive treatment effects is their intuitive interpretation. To see this, write $P(T(s) \in [t, t+a) \mid T(s) \geq t') = \mathbb{E}[1\{T(s) \in [t, t+a)\} \mid T(s) \geq t', X, V]$. The indicator function is a Bernoulli random variable and its distribution is completely determined by its expectation.

Traditionally, the literature has focused on identifying the (additive) effect on the unconditional survival function, that is, $t' = 0$:

$$P(T(s) \in [t, t+a) \mid) - P(T(s') \in [t, t+a)), \tag{1.3.3}$$

see Fredriksson and Johansson (2008), Crépon, Ferracci, Jolivet, and Van den Berg (2009) and Abbring and van den Berg (2005). This approach precludes dynamic selection, see Abbring and van den Berg (2005) for a discussion.[3] Often though it might be of interest to identify the effect of a treatment assigned at a later point in time only for those who actually would receive the treatment. In the labor market example, such a case would arise if a treatment is targeted at longterm unemployed individuals. In the medical example, due to its side effects, a therapy might be targeted only at patients who are at an advanced stage of a disease. For this reason, we consider the general case of conditioning on survival up to a point $t' = t$ for $0 \leq t = s < s' \leq \infty$, that is

$$P(T(t) \in [t, t+a) \mid T(t) \geq t, X, V) - P(T(s') \in [t, t+a) \mid T(s') \geq t, X, V). \tag{1.3.4}$$

We do not impose a parametric form on the distribution of $T(s)$ and we allow for separability and general dependence of observed and unobserved covariates X and V, respectively. The restriction $t = s$ is necessary to "unify" the dynamic selection between treated and untreated, as discussed in the next subsection. By redefining s to be the time to dropout of a treatment, we can analyze the effect of the length s of a treatment on the distribution of $T(s)$.

[3] Abbring and van den Berg (2005) consider a case with conditioning on a positive elapsed spell duration, $t' > 0$, that is, conditioning on $T(s) > t', t' > 0$, and derive bounds for the effect.

There are two limitations we have to consider. First, not specifying the dependence of the distributions of $T(s)$ and the unobservables V makes it impossible to identify the individual treatment effect 1.3.4. The price to pay for the functional form generality is that we have to average V out. Due to dynamic selection, the distribution of the unobservables might 1) be different in the subpopulation of survivors at some point in time $t > 0$ from the distribution in the whole population and 2) differ among different treatment arms. Therefore, it arises the question over which distribution of V to average. Van den Berg, Bozio, and Dias (2014) suggest the following treatment effects:

$$\begin{aligned} &\mathbb{E}[P(T(s) \in [t, t+a) \mid T(s) \geq t, X, V) - \\ &P(T(s') \in [t, t+a) \mid T(s') \geq t, X, V) \mid T(s) \geq t, X] \end{aligned} \tag{1.3.5}$$

$$\begin{aligned} &\mathbb{E}[P(T(s) \in [t, t+a) \mid T(s) \geq t, X, V) - \\ &P(T(s') \in [t, t+a) \mid T(s') \geq t, X, V) \mid T(s') \geq t, X] \end{aligned} \tag{1.3.6}$$

$$\begin{aligned} &\mathbb{E}[P(T(s) \in [t, t+a) \mid T(s) \geq t, X, V) - \\ &P(T(s') \in [t, t+a) \mid T(s') \geq t, X, V) \mid T(s) \geq t, T(s') \geq t, X]. \end{aligned} \tag{1.3.7}$$

The authors refer to those effects as Average Treatment Effect (ATE) on the Treated Survivors, ATE on the nontreated survivors, and ATE on the survivors respectively. We will adapt these definitions to our framework.

A second limitation is that, due to the possibility of static selection, one can observe only the t-compliers with the treatment. This problem has been discussed in the literature on static treatment effects, see Imbens and Angrist (1994). Their solution is to consider only a treatment effect on the subpopulation of compliers. We adapt this restriction to our dynamic concept of compliance. We condition on $S(t) = t$. This restricts the analysis to the subpopulation of t-compliers, that is, to those individuals who would take the treatment at an elapsed duration of t if they were asked to do so. With these considerations, we define the Average Treatment Effect on the Treated Complying Survivors, shortly TE, as

$$\begin{aligned} TE(t, t'a) := \mathbb{E}[&P(T(t) \in [t, t+a) \mid T(t) \geq t, S(t) = t, X, V) - \\ &P(T(t') \in [t, t+a) \mid T(t') \geq t, S(t) = t, X, V) \mid T(t) \geq t, S(t) = t, X]. \end{aligned} \tag{1.3.8}$$

The effects on the nontreated and on the whole population are defined analogously.[4] The positive constant a is chosen such that $a < t' - t$. This restriction insures a comparison of treated with nontreated individuals. Similarly, the treatment effect on the hazard (HTE) is defined as

$$\begin{aligned} HTE(s, s') := \mathbb{E}[&\theta_{T(s)}(t \mid S(t) = t, X, V) - \\ &\theta_{T(s')}(t \mid S(t) = t, X, V) \mid T(s) \geq t, S(t) = t, X]. \end{aligned} \tag{1.3.9}$$

Remark
An alternative treatment effect that can be considered in this framework is a relative effect on the hazard rate at t, $\theta_{T(s)}(t \mid X, V)/\theta_{T(s')}(t \mid X, V)$. Abbring and van den Berg (2005) prove identification of this treatment effect under multiplicative unobserved heterogeneity, that is, under $\theta_{T(s)}(t \mid X, V) = \theta^*_{T(s)}(t \mid X).V$. We do not pursue this approach here.

[4] In fact, they coincide under the assumptions introduced in the next subsection, see proposition 1.3.1.

1.3.2 Identification of dynamic treatment effects

In this section, we show that there exists a function that links the joint distribution of the observables with the treatment effect. As a result, the treatment effect is identified. We derive this function explicitly. Thus, our identification strategy is constructive in the sense that it provide a guidance for estimation. We adopt the following assumptions:

A1 **(Single treatment)** : for any t it holds either $S(t) = t$ or $S(t) = +\infty$.

A2 **(No anticipation)** : For each real $t' \geq t \geq 0$ and each X, V holds

$$\Theta_{T(t')}(t \mid X, V, S(t) = t) = \Theta_{T(\infty)}(t \mid X, V, S(t) = t),$$

where $\Theta_{T(s)}$ is the integrated hazard of $T(s)$. Similarly, we assume "no anticipation" on the set of noncompliers $\{S(t) = \infty\}$, and on all other subpopulations that occur below.

A3 **(Randomization)** : For the instrument Z it holds

$$i) \quad Z \perp\!\!\!\perp \{T(s), S(t)\} \mid X, V \qquad \text{and} \qquad ii) \quad Z \perp\!\!\!\perp V \mid X.$$

A4 **(Consistency)** For all $t, s \in \mathbb{R}_+ \bigcup \{+\infty\}$

i) $Z = t \Rightarrow S(t) = S$

ii) $S = s \Rightarrow T(s) = T$

1. Assumption A1 defines the possible types of noncompliance. Agents are only allowed to choose between being treated at the assigned point in time and being never treated. A1 precludes the type of choices $S(t) = t'$ for some $t' \neq t$ with $t' < \infty$. A1 is compatible with a setup where the treatment is administered at a single point in calender time and agents have no access to an alternative treatment. This setup corresponds to a one-sided noncompliance in the static treatment evaluation literature. Assumptions A1 and A4 imply together that the actual elapsed duration at which the treatment is received, S, can be either equal to Z or to ∞.

2. Assumption A2 states basically that future treatments are not allowed to influence the past. The assumption can be rewritten in the more intuitive way $P(T(t) \geq t \mid X, V, S(t) = t) = P(T(t) \geq t \mid X, V, S(t) = t)$. It implies that the individual probability of a survival up to t is the same for any two future treatments $t', t'', t \leq t', t''$. In a model with forward looking agents, A2 requires that agents either have no knowledge on the point in time of treatment (i.e. they do not anticipate it) or that they do not act upon that knowledge.[5] Technically, jointly with assumption A3, the "no anticipation" assumption is used to ensure equal pretreatment patterns of dynamic selection in the different treatment arms. Thus, it plays a similar role as the sequential randomization assumption. The "no anticipation" assumption has been used in the biostatistics literature in a stronger form and then adopted by the econometric dynamic treatment literature, see Abbring and van den Berg (2003) for a discussion. Although it is a strong assumption, it appears to be plausible in numerous settings in labor market context. First, often the start of

[5] Or that their actions are ineffective.

a training program and the assignment to treatment are dictated by budget and other administrative reasons and appear to the unemployed as random. Those assigned to the treatment might be chosen at random from all eligible unemployed. Moreover, the assignment may occur without a preliminary notice so that the timing is unexpected to the unemployed. An example for such a setting is found in Fitzenberger, Orlanski, Osikominu, and Paul (2013) who analyze the effect of short term training measures in the German unemployment system. Second, the exact content and point in time of implementation of a policy reform are often a subject to persistent debates. The resulting uncertainty might deter agents from building an anticipation about start and content of the reform. In our empirical application, we argue that this argument in fact holds in the case of the French policy reform PARE.

3. Assumption A3 is a randomization assumption. A3 i) implies that once we condition on observables and unobservables, there is no selection into the different treatment assignments. Taken together, i) and ii) imply the conditional independence assumption

$$Z \perp\!\!\!\perp \{T(s), S(t)\} \mid X. \tag{1.3.10}$$

In the labor market example, A3 requires a stable (macro-) economic environment in the period of consideration. Economic structural brakes and mass layoffs might cause a violation of A3. The implication 1.3.10 is testable.

4. The consistency assumption implies that a potential outcome corresponding to a given treatment is observed if the treatment is actually assigned. Another way to write it is $T = T(S), S = S(Z)$. A4 provides the link between potential outcomes and observations and is necessary for identification. It can be related to a structural interpretation of potential outcomes, see Pearl (2000) for a discussion.

In addition to assumptions 1-4, we implicitly assume that all expressions below exist. This amounts to common support assumptions such as $0 \leq P(S = t \mid X, V, Z = t)$. These assumptions imply either that S and Z are discrete or that at least they have a positive probability mass on t and t'. Whether discrete Z and S impose a restriction on the distribution of T depends on the concrete application. In the medical treatment example, a specific therapy might be assigned only at predetermined, common for everybody, elapsed time intervals of the disease, whereas the life or disease duration itself is a continuous variable. In the labor market example, the administrative duration of unemployment is always discrete. Nevertheless, it is usually modeled in the literature as a continuous variable, especially when it is measured on a daily basis. On the other hand, labor market treatments such as training and counselling measures or financial penalties might be designed to come into force only at coarser time intervals. Therefore, it might be practical to model them as discrete variables.

Suppose for the moment that T is observable. The case with right censoring is considered at the end of this subsection. As a motivation for our identification strategy, consider first the following naive candidates for a treatment effect:

$$P(T \in [t, t+a) \mid T \geq t, X, S = t, Z = t) - P(T \in [t, t+a) \mid T \geq t, X, S = \infty, Z = t) \tag{1.3.11}$$

and

$$P(T \in [t, t+a) \mid T \geq t, X, S = t, Z = t) - P(T \in [t, t+a) \mid T \geq t, X, S = t', Z = t'). \tag{1.3.12}$$

Writing 1.3.11 in the form

$$\mathbb{E}[P(T \in [t, t+a) \mid T \geq t, X, S = t, Z = t, V) \mid T \geq t, X, S = t, Z = t] -$$
$$\mathbb{E}[P(T \in [t, t+a) \mid T \geq t, X, S = \infty, Z = t, V) \mid T \geq t, X, S = \infty, Z = t]$$

makes it clear that it compares averages over two different subpopulations of the same cohort: the t-compliers and the t-noncompliers. These two subpopulations might have different distributions of the unobserved heterogeneity V because the treatment status S is a choice variable. As a consequence, it would hold $V \not\perp S \mid T \geq t, X, Z = t$. We will refer to this consequence as static endogeneity or static selection. The (potential) endogeneity arises immediately with the decision to accept or refuse the treatment. As a result, 1.3.11 would capture not only the treatment effect but also the bias from the static selection. We use the naive treatment effect 1.3.11 to analyze the nature of endogeneity. We compare it to our IV estimator to construct a test for exogeneity, see section 1.3.5 for details. The difference between 1.3.11 and the IV estimator is informative about the selection process. A better understanding of the selection might be used to impose more structure on the model. In our empirical application, an estimator of 1.3.11 is shown to underestimate the positive treatment effect. Hence, the control group must contain many quick exits, which sheds light on the reasons for the non-take up of the reform.

The naive treatment effect 1.3.12 compares the average outcome of the t-compliers from the younger cohort $\{Z = t\}$ with the average outcome of the t'-compliers of the older cohort $\{Z = t'\}$. Due to dynamic selection, this comparison amounts to averaging over two potentially different distributions of V. 1.3.12 can be used to shed light on the nature of this dynamic selection process.

Both examples demonstrate the importance and difficulty of the choice of a treatment and a control groups in a setting with static and dynamic selection. Previous studies either preclude static selection by imposing perfect compliance, or preclude dynamic selection through conditioning on $T > 0$ [6]. We propose a strategy that can deal with both types of selection. The intuition for this strategy is as follows. An appealing choice for a treatment group is the set of compliers from the cohort $\{Z = t\}$: consistency links observed outcomes of the treated compliers with the potential outcomes. Suppose for the moment that we observe the potential compliance status at any point in time. Then, one possible control group for the treated t-compliers from cohort $\{Z = t\}$ would be the not yet treated group of t-compliers from the older cohort who survive at least t time units. The intuition behind this choice is the following. If the the unobserved heterogeneity V has the same distribution in the two cohorts at the point in time of inflow, and if these distributions evolve over time in the same way, then V will have the same distribution in the two cohorts at a later pretreatment elapsed duration $t > 0$. The equality of the distributions of V at $t = 0$ is ensured by the randomization assumptions A3 i) and ii). The dynamics is controlled by the "no anticipation" assumption A2. This idea is first developed in Van den Berg, Bozio, and Dias (2014) for the case of perfect compliance. It amounts to a direct comparison of the average outcomes of two cohorts. The method is closely related to the Regression Discontinuity approach developed by Hahn, Todd, and van der Klaauw (2001) but has the main advantage that it can incorporate censoring and deal with dynamic selection. In a first step, we adapt the result of Van den Berg, Bozio, and Dias (2014) to a setting with endogenous compliance.

[6] Instead of conditioning on $T > t$ for some $t > 0$.

Proposition 1.3.1. *Let F be a cdf. Under Assumptions A2 to A4, it holds for all* $\infty \geq t' \geq t \geq 0$

$$F_{V|T(t)\geq t,X,S(t)=t} = F_{V|T(t')\geq t,X,S(t)=t} = F_{V|T\geq t,X,S=t,Z=t}.$$

Proposition 1.3.1 states that the unobservables have the same dynamics for two potential treatments on the set of t-compliers. It also links the distribution of V given a potential treatment to the distribution of V in the subpopulation of observed t-compliers, $\{S = t, Z = t\}$. There are two immediate consequences of 1.3.1. First, the treatment effects on the treated survivors, on the nontreated survivors and on all survivors, respectively, coincide. Second, the following result holds:

Corollary 1.3.2.1.

$$\begin{aligned} TE(t,t',a) &= P(T(t) \in [t,t+a) \mid T(t) \geq t, X, S(t) = t) \\ &- P(T(t') \in [t,t+a) \mid T(t') \geq t, X, S(t) = t) \end{aligned}$$

Corollary 1.3.2.1 provides a direct hint how to choose the treatment group. t-compliers from the cohort $\{Z = t\}$ reveal their preferences at the point in time of treatment. We can therefore link potential and observed outcomes using A4, proposition 1.3.1 and corollary 1.3.2.1. We show in the proof of proposition 1.3.2 that

$$P(T(t) \in [t,t+a) \mid T(t) \geq t, X, S(t) = t) = P(T \in [t,t+a) \mid T \geq t, X, S = t, Z = t). \tag{1.3.13}$$

The main obstacle for constructing a control group is that we do not observe the compliance status of individuals in the older cohort $\{Z = t'\}$ at elapsed duration t. Agents reveal their preferences at the time of treatment. In line with the argumentation above, due to dynamic selection, the subpopulation of t'-compliers differs from the subpopulation of t-compliers in terms of the distribution of V. The key to identification is the observation, that the potential outcome corresponding to a certain treatment is the sum of potential outcomes of compliers and noncompliers, weighted by their proportions:

$$\begin{aligned} &P(T(t') \in [t,t+a) \mid T(t') \geq t, X) = \\ &P(T(t') \in [t,t+a) \mid T(t') \geq t, X, S(t) = t) P(S(t) = t \mid T(t') \geq t, X) + \\ &P(T(t') \in [t,t+a) \mid T(t') \geq t, X, S(t) = \infty) P(S(t) = \infty \mid T(t') \geq t, X), \end{aligned} \tag{1.3.14}$$

or, in a simplified notation,

$$F_0 = F_{C,0} P_C + F_{N,0} P_N, \tag{1.3.15}$$

where the zero indicates the no-treatment case[7], $t' > t + a$, and, with a temporary abuse of notation, C and N denote compliers and noncompliers, respectively. In order to link $F_{C,0} = (F_0 - F_{N,0}P_N)/P_C$ to observables, it is sufficient to express F_0, P_C, P_N and $F_{N,0}$ in terms of observables. Due to the consistency assumption A4, F_0 is equal to $P(T \in [t,t+a) \mid T \geq t, X, Z = t')$.[8]This is an intuitive result. F_0 is the potential outcome when the treatment t' is assigned

[7]The correct expression should be "not yet treated-case". Under the "no anticipation" assumption, however, this distinction does not matter in the interval $[t, t+a)$.

[8]All formal proofs can be found in the appendix.

whereas $P(T \in [t, t+a) \mid T \geq t, X, Z = t')$ is the actual outcome corresponding to the same treatment. Next, due to the assumptions A2 and A3, P_C and P_N do not depend on the future treatment $t' \geq t$. Therefore, it holds $P_C = P(S(t) = t \mid T(t) \geq t, X)$ and $P_N = P(S(t) = \infty \mid T(t) \geq t, X)$. Furthermore, agents from the cohort $\{Z = t\}$ who survive until elapsed duration t reveal their preference at $T = t$. Hence, $P(S(t) = t \mid T(t) \geq t, X)$ and $P(S(t) = \infty \mid T(t) \geq t, X)$ can be linked via the consistency assumption to the observed average outcomes of the compliers and noncompliers of the cohort $\{Z = t\}$. Finally, we make the important observation that t-noncompliers should have the same average potential outcome at $T = t$ when the assigned treatment is either t or $t' > t+a$. This is an implication of the "no anticipation" and consistency assumptions. In both cases they are not treated between t and $t + a$. Therefore, it holds

$$P(T(t') \in [t, t+a) \mid T(t') \geq t, X, S(t) = \infty) = P(T(t) \in [t, t+a) \mid T(t) \geq t, X, S(t) = \infty). \quad (1.3.16)$$

The right-hand side of 1.3.16 can be linked to observables via A4. F_0 is observed directly in the cohort $\{Z = t'\}$ whereas P_C, P_N and $F_{N,0}$ are identified from the cohort $\{Z = t\}$. In the following proposition we state the main result of our paper.

Proposition 1.3.2. *Let $a \leq t' - t$. Under Assumptions A1-A4, $TE(t, t', a)$ is nonparametrically identified and it holds*

$$TE(t, t', a) = \frac{P(T \in [t, t+a) \mid T \geq t, X, Z = t) - P(T \in [t, t+a) \mid T \geq t, X, Z = t')}{P(S = t \mid T \geq t, X, Z = t)} \quad (1.3.17)$$

The proof is in the appendix. Expression 1.3.2 has an intuitive interpretation. It adjusts the difference between the average observed outcomes in the two cohorts by the probability to be a complier. The adjustment takes account of the fact, that any difference between the two cohorts can be caused only by the compliers. The exogenous Z acts as an instrument for the endogenous S. Thus, our result is a dynamic generalization of the standard static LATE literature. Expression 1.3.2 appears to be similar to the static one-sided noncompliance result of Bloom (1984). This resemblance seems natural in a setting where agents are allowed to refuse an assigned treatment but are not able to select into an alternative treatment arm (i.e. choose a different point in time of treatment).

Unlike in the static treatment evaluation models, randomization alone is not enough to ensure identification. An experiment might be randomized at $t = 0$ but due to dynamic selection endogeneity arises over time. The "no anticipation" assumption precludes this possibility. As a result, our model deals successfully with both static and dynamic selection and thus it provides the link between the standard LATE literature and the literature on dynamic treatment evaluation.

Remark

A special case of 1.3.2 is the limit case $a \to \infty$. We devote a separate section on its identification and estimation because of the importance and specifics of hazards.

Remark 2

An implicit consequence of the assumptions A1-A4 is that the treatment effects do not depend on t' as long as $t' > t$. Therefore, we omit the dependence on t' and write $TE(t, a)$ and $HTE(t)$.

A comparison to related methods should highlight the advantages of our approach and the tradeoff between the strength of assumptions and the generality of results. The recent literature on dynamic treatment evaluation can be roughly divided into (dynamic) matching estimators, dynamic discrete choice models and the Timing-of-Events approach.[9] Some influential studies in the first category are the papers of Sianesi (2004), Fredriksson and Johansson (2008) (henceforth abbreviated as FJ) and Crépon, Ferracci, Jolivet, and Van den Berg (2009) (CFJB). All of them achieve nonparametric identification of an additive average treatment effect.[10] Sianesi (2004) assumes that there is no unobserved heterogeneity and adapts a conditional independence assumption (CIA). Similarly, FJ assume that the unobserved heterogeneity does not jointly determine the employment and treatment assignment. Thus, both methods require very rich data sets. Additionally, FJ do not develop a method for constructing the standard errors. CFJB adopt a CIA and a "no anticipation" assumption. All three papers assume perfect compliance.

Next, some important contributions in the literature on dynamic discrete choice models are made for example in Taber (2000) and Heckman and Navarro (2007)(HN). See Abbring and Heckman (2007) for additional references. These approaches typically heavily rely on the "identification at infinity" approach, which assumes that the support of the exclusion restriction is very large. We discuss only HN here (its reduced form model). The biggest advantage of their approach is the complex nature of the identified treatment effects. First, they are able to consider potential outcomes $Y(s,t)$, $t \geq s$, where t is the point in time of realization of the outcome (such as e.g. age of the agent) and s is the point in time at which the treatment is received. We only consider the case $t = s$.[11] This is not a serious drawback of our approach though: by redefining S to be the duration of a stay in a certain treatment, we are also able to analyze the effect of the duration of the treatment. More importantly, HN are able to identify the joint distributions of the counterfactuals. This is an important advantage as it allows answering more complex policy questions. Furthermore, HN's approach allows for a very general time-varying nature of the unobservables. Their identification is based on an explicit modelling of the selection, as well as on a factor structure of the unobservables. There are several potential restrictions of their approach. First, the explicit modelling of selection requires large variation in the exclusion restriction ("identification at infinity"). Moreover, their identification strategy crucially depends on the separability and independence of observed and unobserved covariates. We allow X and V to be arbitrary dependent and need no separability assumptions. Next, the price to pay for the identification of joint distributions is the factor analysis assumptions. Normalization and exclusion restrictions are rather arbitrary, unless a measurement system for the choice equations (proxies for the factors) is available. In addition, and maybe most importantly, the approach of HN considers only perfect compliance.

The last category of methods, the Timing-of-Events approach, is developed in the influential paper of Abbring and van den Berg (2003). Similarly to HN, the main advantage of this method is the generality of the treatment effect that can be identified. Identifications relies on the semiparametric structure of the Mixed Proportional Hazard (MPH) method and on the independence of unobserved and observed covariates. As in the previous papers discussed

[9] See Abbring and Heckman (2007) for a survey.

[10] A direct effect, $\mathbb{E}[Y_1 - Y_0]$, or a distributional one.

[11] In particular, in the hazard case $a \to 0$ this amounts to considering only an instantaneous treatment effect.

here, noncompliance is not considered.

Thus far, we have assumed we can observe the whole length of spells in the state of interest, T. A typical feature of duration data is that observations might be censored. In this paper, we consider right censoring.[12] In labor market studies, right censoring typically arises when at the end of the study the individuals are still unemployed, so the unemployment spell has an unknown length. The unemployed might also simply stop attending the training and drop out of the study (sample attrition). In addition, the job search might be interrupted by a transition out of the labor force due to maternity, sickness, military service or other reasons. In biomedical studies, and particularly in clinical trials, spells are right-censored when patients die from another cause (competing risks) or withdraw from treatment. We introduce formally right censoring in the following way: let C be a real nonnegative random variable. We observe $(\widetilde{T}, \delta)$ and not directly (T, C). It is not possible to recover nonparametrically the joint distribution of T and C from the distribution of $(\widetilde{T}, \delta)$ without additional assumptions. The reason for this impossibility is a nonidentification result that goes back to Cox (1962) and Tsiatis (1975), namely that to each pair of latent variables (T_d, C_d) there exists an independent pair of variables (T_i, C_i) that is observationally equivalent to (T_d, C_d). To achieve identification, we adopt the following additional standard assumption:

A5) (Random censoring)

$$C \perp\!\!\!\perp (T, S) \mid X, Z.$$

Assumption A5 is nontestable due to the nonidentification result of Tsiatis (1975). In the context of our empirical application, we show with a Monte Carlo simulation that plausible violations of A5 offset each other. Thus, our estimation results are likely to be robust against violations of A5. With A5, we can prove the following proposition:

Proposition 1.3.3. *Under assumptions **A1 - A5** TE(t,X) is identified.*

The proof of 1.3.3 is straightforward. The probabilities $P(T \in [t, t+a) \mid T \geq t, X, Z = j)$ for $j \in \{t, t'\}$ can be written as differences of survival functions and be estimated consistently with a Kaplan-Meier estimator, see section 1.3.3 for details. Note also that S is observed whenever $\widetilde{T} \geq t$, so that due to A5

$$P(S = t \mid T \geq t, X, Z = t) = P(S = t \mid T \geq t, C \geq t, X, Z = t) = P(S = t \mid \widetilde{T} \geq t, X, Z = t). \quad (1.3.18)$$

The last probability in 1.3.18 contains only observables and can be consistently estimated from the data.

1.3.3 IV estimation of dynamic treatment effects

To ease notation, probability and survival functions concerning the cohorts $\{Z = t\}$ and $\{Z = t'\}$ are denoted with an index 1 and 2, respectively. For example, we write $P_1(T \in [t, t+a) \mid T \geq t, Z = t)$ instead of $P(T \in [t, t+a) \mid T \geq t, Z = t)$. Furthermore, we ignore the dependence on observed covariates X. Assumptions A2 and A3 are adapted accordingly. The generalization to the case with covariates is straightforward. Denote with $\bar{F}_1$ and $\bar{F}_2$

[12] Extensions to left or interval censoring are straightforward.

the survival functions of T in the two cohorts, $\bar{F}_i(t) := P_i(T > t)$. A starting point for our estimation procedure is the equality

$$TE(t,a) = \frac{1}{P_1(S = t \mid T \geq t)}\Big(\frac{\bar{F}_2(t+a)}{\bar{F}_2(t)} - \frac{\bar{F}_1(t+a)}{\bar{F}_1(t)}\Big), \tag{1.3.19}$$

which holds under assumptions A1-A4. It follows from the result in proposition 1.3.2 together with $P_i(T \in [t, t+at) \mid T \geq t) = 1 - \bar{F}_i(t+a)/\bar{F}_i(t)$. T might be censored so that we only observe $(\widetilde{T}, \delta)$. $\bar{F}_i(t)$ can be consistently estimated with the Kaplan-Meier estimator. Under the independent censoring assumption A5 and additional mild regularity conditions, it holds

$$\widehat{\bar{F}}_i(t) = \bar{F}_i(t) + o_p \qquad \text{and} \tag{1.3.20}$$

$$\sqrt{n}\left(\widehat{\bar{F}}_i(t) - \bar{F}_i(t)\right) \xrightarrow{d} N(0, \sigma_i(t)) \qquad \text{as} \quad n \to \infty, \tag{1.3.21}$$

where $\sigma_i(t)$ is the asymptotic variance of the Kaplan-Meier estimator. The additional regularity conditions can be found in standard references for survival analysis, see e.g. Andersen, Borgan, Gill, and Keiding (1997), chapter IV.3 or Kalbfleisch and Prentice (2002), chapter 5.6. We refer to them as KM conditions and do not state them explicitly (all results hold for both continuos and discrete time).

Next, under the independent censoring assumption, it holds

$$P_1(S = t \mid T \geq t) = P(S = t \mid T \geq t, Z = t, C \geq t) = P_1(S = t \mid \widetilde{T} \geq t) =: p > 0. \tag{1.3.22}$$

p contains only observables and is nonparametrically identified. Let $\widehat{p} := \widehat{P_1}(S = t \mid \widetilde{T} \geq t)$ be a consistent nonparametric estimator of p. We define the IV-estimator $\widehat{TE}(t,a)$ of $TE(t,a)$ as

$$\widehat{TE}(t,a) = \frac{1}{\widehat{p}}\Big(\frac{\widehat{\bar{F}_2}(t+a)}{\widehat{\bar{F}_2}(t)} - \frac{\widehat{\bar{F}_1}(t+a)}{\widehat{\bar{F}_1}(t)}\Big). \tag{1.3.23}$$

Its properties follow from the properties of the Kaplan-Meier estimator. The following proposition states the consistency of 1.3.23.

Proposition 1.3.3.1. *Under assumptions A1-A5 and the KM conditions, it holds*

$$\widehat{TE}(t,a) - TE(t,a) = o_p$$

for each admissible pair (t,a).

This result follows directly from the continuity of the function $G(a,b,c,d,e) = \frac{1}{e}\left(\frac{a}{b} - \frac{c}{d}\right)$, the Continuous Mapping Theorem and the consistency of $\bar{F}_i(t)$ and $\widehat{p}$.

Consider the Null hypothesis

$$H_0: \qquad \text{(Ineffective treatment)} \qquad \frac{\bar{F}_2(t+a)}{\bar{F}_2(t)} - \frac{\bar{F}_1(t+a)}{\bar{F}_1(t)} = 0. \tag{1.3.24}$$

Under 1.3.24, it holds

$$\begin{aligned}\sqrt{n}\widehat{TE}(t,a) &= \frac{\sqrt{n}}{\widehat{p}}\Big(\frac{\widehat{\bar{F}_2}(t+a)}{\widehat{\bar{F}_2}(t)} - \frac{\widehat{\bar{F}_1}(t+a)}{\widehat{\bar{F}_1}(t)}\Big) = \\ &= \frac{\sqrt{n}}{\widehat{p}}\Big(\frac{\widehat{\bar{F}_2}(t+a)}{\widehat{\bar{F}_2}(t)} - \frac{\bar{F}_2(t+a)}{\bar{F}_2(t)}\Big) - \frac{\sqrt{n}}{\widehat{p}}\Big(\frac{\widehat{\bar{F}_1}(t+a)}{\widehat{\bar{F}_1}(t)} - \frac{\bar{F}_1(t+a)}{\bar{F}_1(t)}\Big)\end{aligned}$$

For $i = 1,2$ the Taylor expansion of $\frac{\widehat{\bar{F}}_i(t+a)}{\widehat{\bar{F}}_i(t)}$ around $\frac{\bar{F}_i(t+a)}{\bar{F}_i(t)}$ can be written as

$$\begin{aligned}&\frac{\widehat{\bar{F}}_i(t+a)}{\widehat{\bar{F}}_i(t)} = \frac{\bar{F}_i(t+a)}{\bar{F}_i(t)} + \frac{1}{\bar{F}_i(t)}(\widehat{\bar{F}}_i(t+a) - \bar{F}_i(t+a)) - \frac{\bar{F}_i(t+a)}{\bar{F}_i^2(t)}(\widehat{\bar{F}}_i(t) - \bar{F}_i(t)) \\ &+O\Big[(\widehat{\bar{F}}_i(t+a) - \bar{F}_i(t+a))(\widehat{\bar{F}}_i(t) - \bar{F}_i(t)) + (\widehat{\bar{F}}_i(t) - \bar{F}_i(t))^2\Big],\end{aligned}$$

and therefore

$$\begin{aligned}&\sqrt{n}\Big(\frac{\widehat{\bar{F}}_i(t+a)}{\widehat{\bar{F}}_i(t)} - \frac{\bar{F}_i(t+a)}{\bar{F}_i(t)}\Big) = \frac{\sqrt{n}}{\bar{F}_i(t)}(\widehat{\bar{F}}_i(t+a) - \bar{F}_i(t+a)) - \frac{\bar{F}_i(t+a)\sqrt{n}}{\bar{F}_i^2(t)}(\widehat{\bar{F}}_i(t) \\ &-\bar{F}_i(t)) + O\Big[\sqrt{n}(\widehat{\bar{F}}_i(t+a) - \bar{F}_i(t+a))(\widehat{\bar{F}}_i(t) - \bar{F}_i(t)) + \sqrt{n}(\widehat{\bar{F}}_i(t) - \bar{F}_i(t))^2\Big].\end{aligned}$$

The last term converges to zero in probability.

With 1.3.21, the terms $\frac{\sqrt{n}}{S\bar{F}_i(t)}(\widehat{\bar{F}}_i(t+a) - \bar{F}_i(t+a))$ and $\frac{\bar{F}_i(t+a)\sqrt{n}}{\bar{F}_i^2(t)}(\widehat{\bar{F}}_i(t) - \bar{F}_i(t))$

are asymptotically normally distributed with mean 0 and variances

$$\frac{1}{\bar{F}_i^2(t)}\sigma_i(t+a) \quad \text{and} \quad \frac{\bar{F}_i^2(t+a)}{\bar{F}_i^4(t)}\sigma_i(t), \quad \text{respectively.}$$

With the independence of the random variables D_1 and D_2, where $D_i = \frac{\widehat{\bar{F}}_i(t+a)}{\widehat{\bar{F}}_i(t)}$, $i = 1,2$, we can now state the following proposition.

Proposition 1.3.3.2. *Let assumptions A1-A5 and the KM conditions hold. Then, under the null 1.3.24, it holds*

$$\widehat{TE}(t,a) \xrightarrow{d} N\Big(0, \frac{1}{p^2}\sum_{i=1}^{2}\Big(\frac{1}{S_i^2(t)}\sigma_i(t+a) + \frac{\bar{F}_i^2(t+a)}{\bar{F}_i^4(t)}\sigma_i(t) + \frac{\bar{F}_i(t+a)}{\bar{F}_i^3(t)}\sigma_i(t,t+a)\Big)\Big), \tag{1.3.25}$$

where $\sigma_i(t,t+a)$ is the covariance of $\widehat{\bar{F}}_i(t)$ and $\widehat{\bar{F}}_i(t+a)$.

Confidence bands can be constructed by replacing the unknown terms in the variance with consistent estimates, for example using the Greenwood's formula, see Andersen, Borgan, Gill, and Keiding (1997). It follows from 1.3.25 that the precision of the estimator is inversely related to p. The bigger the compliance probability p, i.e. the stronger the instrument Z for the endogenous S, the smaller the variance of the IV-estimator. This intuitive result is in line with the standard static IV literature. 1.3.23 can be interpreted as a dynamic version of the Wald estimator. A generalization to the case of covariates can be achieved by replacing the unconditional Kaplan-Meier estimator with the conditional estimator of Gonzalez-Manteiga and Cadarso-Suarez (2007), following the same steps as here.

1.3.4 Identification and estimation of additive treatment effects on the hazard

In this subsection, we state conditions under which the treatment effect on the hazard, 1.3.9, is identified and develop the estimation theory. The HTE deserves a special attention for two reasons. First, the hazard of the duration variable represents the most interesting feature of its distribution in multiple applications, see Van den Berg (2001) for various examples and a discussion. Second, estimation of hazard effects in a treatment evaluation framework involves estimation at the boundary of the admissible domain. We develop an estimator that takes into account the region of estimation and does not lead to an increased bias.

Identification

Write $W = (X, V)$ and let Ω_W be the set of possible values for W. Further, write $\Psi(t \mid X) := HTE(t, X)$ (we stress explicitly the dependence on X) and define $\theta(t \mid X) := \lim_{dt \to 0} P(T \in [t, t+dt \mid T \geq t, X))/dt$ (all expressions are assume to exist). The rest of the notation is the same as in the last sections. Again we assume access to an i.i.d. sample

$$(\widetilde{T}_1, S_1, Z_1, X_1, \delta_1), \ldots, (\widetilde{T}_n, S_n, Z_n, X_n, \delta_n).$$

Our first result is the following

Proposition 1.3.4. *Let the measurable function $g : \Omega_W \to \mathbb{R}^+$ fulfill $\mathbb{E}[g(W)] < \infty$ and $\mid \theta(t \mid w) \mid \leq g(w)$ for each $w \in \Omega_W$. Then, under assumptions A1-A5, $\Psi(t \mid X)$ is identified and it holds*

$$\Psi(t \mid X) := \frac{\theta(t \mid X, Z = t) - \theta(t \mid X, Z = t')}{P(S = t \mid T \geq t, X, Z = t)}. \tag{1.3.26}$$

Under the Lebesque dominated convergence theorem,

$$\theta(t \mid X) = \lim_{dt \to 0} \mathbb{E}[P(T \in [t, t+dt) \mid T \geq t, X, V)/dt \mid T \geq t, X] = \mathbb{E}[\theta(t \mid X, V)],$$

and the proof follows directly from proposition 1.3.2. Thus, as expected, the *HTE* is revealed to be the limit case of the general treatment effect *TE*, $HTE = lim_{dt \to 0} TE/dt$. In the case of a full compliance, that is $P(S = t \mid T \leq t, X, Z = t) = 1$, *HTE* reduces to $\theta(t \mid X, Z = t) - \theta(t \mid X, Z = t')$ which is the result of Van den Berg, Bozio, and Dias (2014).

Estimation

Henceforth, we denote with $\theta_1(t \mid X)$ the hazard $\theta(t \mid X, Z = t$ of the younger cohort , $\{Z = t\}$, and with $\theta_2(t \mid X)$ the hazard $\theta(t \mid X, Z = t')$ of the older cohort. If the treatment is effective, there will be a structural break in the hazard at the moment of treatment. Hence, when estimating $\Psi(t \mid X)$, only the observations $\widetilde{T}$ that are bigger than or equal to t are informative about $\theta_1(t \mid X)$.[13]This leads to estimating a hazard at the left boundary of the interval $[t, \bar{T})$ where $\bar{T}$ is some maximum duration, possibly ∞. Smooth hazard estimators that use a symmetric kernel would have a large bias at t, a problem called boundary effect in the

[13]This does not apply to $\theta_2(t \mid X)$.

literature, Müller and Wang (1994). Without loss of generality, let $[0,1]$ be the set of possible values of the duration variable and $b = b(n)$ a bandwidth of a kernel estimator, $b < 0.5$. The set $B_L := \{t : 0 \leq t < b\}$ is called a left boundary region (we do not discuss problems arising at the right boundary here). Employing a symmetric kernel to estimate the hazard at a point from that region could lead to a high bias, because the support of the kernel exceeds the range of the data. In the interior $(0,1)$, this is only a finite sample problem. At the boundary $t = 0$, the problem persists with increasing sample size n. Boundary problems are not endemic to hazards, they arise also in the estimation of a density function, see Karunamuni and Alberts (2005). Müller and Wang (1994) develop a class of asymmetric kernels and use them to adapt the unconditional Ramlau-Hansen estimator to the boundary case. The kernels vary with the point of estimation and have a support that does not exceed the range of the duration variable. These kernels are referred to as boundary kernels. Following this approach, we adapt the conditional kernel hazard estimator of Nielsen and Linton (1995) to the case of estimation at the boundary by using boundary kernels. For simplicity, we assume that we estimate $\Psi(t \mid x)$ at an interior point x of Ω_X. Let k be a symmetric one-dimensional density function with support $[-1,1]$, that is

$$\int_{-1}^{1} k(y)dy = 1 \qquad \text{and} \qquad \int_{-1}^{1} yk(y)dy = 0$$

and define k_1 and k_2 as

$$k_1 = \int_{-1}^{1} y^2 k(y)dy \qquad \text{and} \qquad k_2 = \int_{-1}^{1} k^2(y)dy.$$

Define the q-dimensional product kernel $K(x) = \Pi_{i=1}^{q} k(x(i))$, where $x = (x(1), \ldots, x(q))$. Next, let k_+ denote the asymmetric kernel function

$$\begin{aligned} &k_+ : [0,1] \times [-1,1] \to \mathbb{R} \\ &(h,y) \to \frac{12}{(1+h)^4}(y+1)[y(1-2h) + (3h^2 - 2h + 1)/2]. \end{aligned}$$

This is a boundary kernel function as defined in Müller and Wang (1994). The support of $k_+(h,.)$ is $[-1,h]$. In analogy to the symmetric kernel k, we define the second moments of $k_+(0,.)$ as

$$k_1^+ = \int_{-1}^{0} y^2 k_+(0,y)dy \qquad \text{and} \qquad k_2^+ = \int_{-1}^{0} k_+^2(0,y)dy.$$

Using standard counting processes notation, define for $i = 1, \ldots, n$ the observed failure process of the i$^{\text{th}}$ individual at time t, $N_i(t) := 1\{\widetilde{T}_i \leq t, T_i \leq C_i\}$ and the individual process at risk, $Y_i(t) := 1\{\widetilde{T}_i \geq t\}$. To differentiate between observations from the cohorts 1, that is $\{Z = t\}$, and 2, that is $\{Z = t'\}$, we add a subscript 1 or 2, respectively. For example, $X_{1,i}$ denotes an observation of X that comes from the cohort $\{Z = t\}$. Then our estimator $\widehat{\Psi}(t \mid x)$ of $\Psi(t \mid x)$ is defined as

$$\begin{aligned} \widehat{\Psi}(t \mid x) := \frac{1}{\widehat{p_1(t \mid x)}} \Big(&\frac{\sum_{i=1}^{n} K(\frac{x - X_{1,i}}{b}) \int k_+(t, \frac{t-s}{b}) dN_{1,i}(s)}{\sum_{i=1}^{n} K(\frac{x - X_{1,i}}{b}) \int k_+(t, \frac{t-s}{b}) Y_{1,i}(s)ds} \\ &- \frac{\sum_{i=1}^{n} K(\frac{x - X_{2,i}}{b}) \int k_+(t, \frac{t-s}{b}) dN_{2,i}(s)}{\sum_{i=1}^{n} K(\frac{x - X_{2,i}}{b}) \int k_+(t, \frac{t-s}{b}) Y_{2,i}(s)ds} \Big), \end{aligned} \tag{1.3.27}$$

where $\widehat{p_1(t \mid x)}$ is a consistent nonparametric estimator for $p_1(t \mid x) := P(S = t \mid T \geq t, X = x, Z = t)$. The term

$$\widehat{\theta_j(t \mid x)} := \frac{\sum_{i=1}^n K(\frac{x-X_{j,i}}{b}) \int k_+(t, \frac{t-s}{b}) dN_{j,i}(s)}{\sum_{i=1}^n K(\frac{x-X_{j,i}}{b}) \int k_+(t, \frac{t-s}{b}) Y_{j,i}(s) ds}$$

for $j = 1, 2$ is a conditional smooth hazard estimator for $\theta_j(t \mid x)$ developed in Nielsen and Linton (1995) and adapted to the boundary case. Define

$$\theta_j^*(t \mid x) := \frac{\sum_{i=1}^n K(\frac{x-X_{j,i}}{b}) \int k_+(t, \frac{t-s}{b}) \theta_j(s \mid X_{j,i}) Y_{j,i}(s) ds}{\sum_{i=1}^n K(\frac{x-X_{j,i}}{b}) \int k_+(t, \frac{t-s}{b}) Y_{j,i}(s) ds} \qquad j = 1, 2 \tag{1.3.28}$$

and

$$\Psi^*(t \mid x) = \frac{1}{\widehat{p_1(t \mid x)}} (\theta_1^*(t \mid x) - \theta_2^*(t \mid x)). \tag{1.3.29}$$

We need the following assumptions.

H1 $\mathbb{E}[Y_i(s)] = u(s)$ and $u(.)$ is continuous

H2 i) $f(t,x)u(t)$ is positive on a neighbourhood U of $(0, x_0)$, where x_0 is an interior point of Ω_X and f is the density function. ii) θ_j is twice continuously differentiable on U iii) fu is continuously differentiable on U.

H3 $nb^{q+1} \to \infty$ and $b = b(n) \to \infty$ as $n \to \infty$.

The following proposition states the pointwise asymptotic properties of $\widehat{\Psi}(0 \mid x_0)$.

Proposition 1.3.5. *Under assumptions H1-H3, the following results hold:*

i) $\sqrt{nb^{q+1}}(\widehat{\Psi}(0 \mid x_0) - \Psi^*(0 \mid x_0)) \xrightarrow{d} N\big[0, k_2^+ k_2^q \frac{1}{p_1^2(0|x_0)} (\theta_1(0 \mid x_0)/f_1(0,x_0) + \theta_2(0 \mid x_0)/f_2(0,x_0))\big]$

iii) $\frac{nb^{q+1}}{\widehat{p_1(0|x_0)}^2} \sum_{j=1}^2 \frac{\sum_{i=1}^n K^2(\frac{x_0-X_{j,i}}{b}) \int k_+^2(\frac{-s}{b}) dN_{j,i}(s)}{(\sum_{i=1}^n K(\frac{x_0-X_{j,i}}{b}) \int k_+(\frac{-s}{b}) Y_{j,i}(s) ds)^2} \xrightarrow{p} k_2^+ k_2^q \frac{1}{p_1^2(0|x_0)} (\theta_1(0 \mid x_0)/f_1(0,x_0) + \theta_2(0 \mid x_0)/f_2(0,x_0))$

Result i) gives the asymptotic distribution of the estimator, ii) characterizes the bias and iii) provides the standard errors for confidence bounds around Ψ^*. If the bandwidth is chosen to be of $o(n^{-1/(q+5)})$, then the asymptotic bias is negligible and proposition 1.3.5 can be used to construct confidence bands for Ψ.

1.3.5 Framework for the analysis of endogeneity

Understanding the nature of selection is important for setting up and evaluating a policy reform. Often a comprehensive policy reform is preceded by a small scale pilot study that allows for noncompliance. Understanding the non-take up of the pilot study might help better design the reform and derive bounds for its effect under perfect compliance. Better understanding of the endogeneity reasons can be used to model explicitly the selection process in more complex (e.g. general equilibrium) models. We develop a framework for answering the following two questions:

i) Is there endogenous selection caused by the decision of the agents to accept or refuse the treatment?

i)) If yes, in which direction would be the bias caused by the endogenous selection?

Answering the first question requires a specification of the possible channels of (static) endogeneity. In our framework, there are two potential endogeneity channels. First, unobserved characteristics of the agents determine both potential outcomes and the potential compliance decision. Second, the potential outcome itself (that is, after "controlling" for observed and unobserved individual characteristics) might influence the potential compliance status. The first channel amounts to a violation of

$$S(t) \perp\!\!\!\perp \{T(s)\} \mid X \tag{1.3.30}$$

and the second of

$$S(t) \perp\!\!\!\perp \{T(s)\} \mid X, V \tag{1.3.31}$$

We preclude the possibility of a violation of 1.3.31: we assume that the only way the potential outcome might influence the decision $S(t)$ is that the agent might have a knowledge of $T(s)$ and use it in the decision process. This individual knowledge of the potential outcome (or its distribution) is unobserved by the econometrician. It is therefore included in V.[14] With these considerations, we define the following null hypothesis:

$$H_0 : \qquad S(t) \perp\!\!\!\perp \{T(s)\} \mid X \tag{1.3.32}$$

For Borel-measurable sets B, 1.3.32 implies the following relation:

$$\widetilde{H}_0 : P(T(\infty) \in B \mid T(\infty) \geq t, X, S(t) = t) - P(T(\infty) \in B \mid T(\infty) \geq t, X, S(t) = \infty) = 0. \tag{1.3.33}$$

Using A1-A4 and following the steps in proof of proposition 1.3.2, we obtain the equivalent relation

$$\widetilde{H}_0 : \qquad P(T \in B \mid T \geq t, X, Z = t') - P(T \in B \mid T \geq t, X, S = \infty, Z = t) = 0. \tag{1.3.34}$$

Intuitively, if there is no selection, then the average observed outcomes of nontreated compliers and noncompliers should be the same. As a result, the average observed outcome of the whole cohort $\{Z = t'\}$ under no treatment (the left-hand side of 1.3.34) should be equal to the average observed outcome of the noncompliers from the cohort $\{Z = t\}$ (the right-hand side of 1.3.34). Equation 1.3.34 contains only observables. Deriving a distribution of a test statistics in the case $B = [t, t + a)$ follows precisely the same steps as for the null hypothesis 1.3.24. We omit it here. A simplified testing procedure would induce a comparison of survival functions. The corresponding null hypothesis is

$$\widetilde{H}_0 : \qquad P(T \geq t \mid X, Z = t') - P(T \geq t \mid X, S = \infty, Z = t) = 0. \tag{1.3.35}$$

A test statistics is constructed by replacing the theoretical probabilities with their Kaplan-Meier estimators.

[14] We have to assume that the agent does not learn about the potential outcomes over time. With a time-varying V, we would lose identification.

To answer question ii), we can compare the (theoretical proper) treatment effect 1.3.24 to the naive treatment effect 1.3.11. Written in the simplified notation of section 1.3.2, this is a comparison of a) $F_{C,1} - F_{C,0}$ and b) $F_{C,1} - F_{N,0}$. This ad hoc approach can be justified with 1.3.33. Recall that $F_{C,0} = (F_0 - F_{N,0}P_N)/P_C$. Subtracting b) from a), we obtain

$$(F_{C,1} - (F_0 - F_{N,0}P_N)/P_C) - (F_{C,1} - F_{N,0}) =$$
$$(F_{N,0}P_C + F_{0,N}P_N - F_0)/P_C = (F_{N,0} - F_0)/P_C.$$

The nominator $(F_{N,0} - F_0)$ of the last expression is precisely the left-hand side of 1.3.33.

1.4 Empirical Application: the French PARE labour market reform from 2001

1.4.1 Research question and description of the reform

We combine the IV method we developed in section 1.3 with a unique empirical strategy to analyze the effect of a reform in the French unemployment insurance system on the duration of unemployment. The new system, called Plan d'Aide au Retour á l'Emploi (PARE hereafter), brings about two main changes. First, the insurance benefit digression is abolished. Under the old system, called Allocation Unique Degressive (AUD), the size of the payments depends on the elapsed duration of unemployment and decreases stepwise at the end of predefined intervals. Under the new system, benefits remain at a fixed level for the whole payment period. Second, the new system introduces a variety of ALMP measures. The first one is compulsory meetings on a regular basis with a caseworker. During the first meeting, a personal plan called Plan d'Action Personalisé (PAP) is established. It captures in a contract the details about the degree of assistance provided by the caseworker to the unemployed as well as the targeted job type and the region of search. This contract is updated periodically if the individual remains unemployed, typically every six months. During the first meeting the unemployed is also assigned to one of different types of services such as counseling and training, see Freyssinet (2002) for a detailed description of the reform.

The PARE reform has two unique characteristics. First, individuals whose unemployment spells started before the implementation of the reform and were still unemployed during its commencement were given the option to choose whether they want to stay in the AUD regime or switch to PARE. If an unemployed decides to stay in AUD, his benefits payments remain in the digression scheme and no further changes of the status quo take place. If an unemployed decides to switch to PARE, his benefit payments are fixed at the last level payed and no further digression occurs until the end of the payment period (or unemployment exit).[15] This option does not apply to spells starting after the 1st of July 2001, the day of coming into force of PARE. All new unemployed are automatically assigned to the new system.

Second, the new system is unique in terms of its generosity. Although the meetings with the case worker were mandatory, there was no actual monitoring of the job search efforts.

[15]The individuals indicate their decision per mail.

Furthermore, the individuals could generally refuse to take part in assigned training or counseling measures without incurring any sanctions. Thus, the better financial conditions of PARE were not linked to a real reinforcement of ALMP measures.

Ex ante it is not clear what the overall effect of the reform would be. On the one hand, abolishing the digression of benefits removes an incentive for a high search effort. Therefore it can be expected that the exit rate from unemployment to employment will decrease. This intuition is incorporated in theoretical models on optimal unemployment insurance design, see e. g. Pavoni and Violante (2007). There is also some empirical evidence for it, for example in Prieto (2000) and Dormont, Fougère, and Prieto (2001). These papers use a parametric specification (of the PH and MPH models, respectively) to compare the French unemployment insurance system from 1986-1992, which is characterized by a single drop in benefits, with its successor, the digressive AUD system. Dormont, Fougère, and Prieto (2001) conclude that under AUD the return to employment is slowed down. Further insights about the influence of the structure of the insurance system on the unemployment dynamics can be gained when a change from a system with several drops in the entitlement to a system with a single drop is interpreted as a combination of several changes in the potential benefit duration and the entitlement amount. Papers on the influence of the potential duration of unemployment benefits generally find that prolonged benefit duration increases the duration of unemployment, see Lalive (2008) and Lalive, van Ours, and Zweimüller (2006) for studies with Austrian data and Katz and Meyer (1990) for a study with US data. Similarly, studies on the effect of the amount of the benefit entitlement find that a benefit reduction has a positive effect on the exit rate out of unemployment, see for example Lalive, Zweimüller, and van Ours (2005) for a study with Swiss data and Lalive, van Ours, and Zweimüller (2006) for a study with Austrian data.

On the other hand, active labor market policies are supposed to enhance the job search and increase the exit rate to employment. A vast body of empirical literature investigates the effects of training, counseling and subsidized wages on the employment dynamics, see Heckman, LaLonde, and Smith (1999) and Kluve (2010) for an overview. The results of the training literature are quite heterogeneous. While in a study of Swedish policy Richardson and den Berg (2001) find significant positive effects of a vocational training on the transition rate to work, Gritz (1997) finds that a participation in a private training program in the US can induce very different results across genders, and Crépon, Ferracci, and Fougère (2007) and Crépon, Ferracci, Jolivet, and Van den Berg (2009) find little or no effect in studies with French labor market data. Similarly, Crépon, Dejemeppe, and Gurgand (2005) and Van den Berg, Kjærsgaard, and Rosholm (2012) find a significantly positive impact of counseling on the exit rate to unemployment, while Van den Berg and Van der Klaauw (2010) establish at best a small effect. Wage subsidies on the contrary are shown generally to have beneficial influence, see Blundell, Dias, Meghir, and Reenen (2004), Gerfin and Lechner (2002) and Kluve (2010). Additionally to these direct of effects, the thread effects of ALMP should be taken into account. Such effects can be positive, see e. g. Rosholm and Svarer (2008), or negative, Crépon, Ferracci, Jolivet, and Van den Berg (2010).

1.4.2 The Data

The data sample we use is taken from a matching of two administrative data sets: the Fichier Historique (FH) data set, which contains information about the unemployment spells and is issued by the French public employment agency (Agence nationale pour l'emploi, ANPE), and the Déclaration Anuelle de Données Sociales (DADS) data set, which contains the employment information of all individuals employed in the private sector and is issued by the French Statistical Institute (Insee). We extract a set of variables, rich enough to account for the socio-economic status of the individuals , namely age, gender, marital status, number of children, educational level, professional experience, description of the job position/type in the last employment spell, reason for entering unemployment, exit direction (out of unemployment), and unemployment history. Details about the construction and content of the variables can be found in appendix 1.6.2.

To preclude geographical heterogeneity we restrict our sample to the administrative region Îll de France, which contains Paris and consists of the administrative departments 75, 77, 78, 91, 92, 93, 94 and 95. Because of its size and specific infrastructure, this region might differ from the rest of France in terms of labor market dynamics (mobility, unemployment structure, wages) and in terms of the implementation of the reform. Moreover, the macroeconomic conditions in this region are stable over the period of consideration, which insures the comparability of the cohorts, see subsection 1.4.4.

The choice of the cohorts is restricted by the available data. There is no administrative variable that captures the compliance status of the unemployed. Moreover, due to budget reasons, there was a considerable time variation in giving the treatment across individuals. Thus, some individuals might have exited the state of unemployment before receiving their assigned treatment. We develop a novel empirical strategy to deal with this problem. We choose the younger cohort $\{Z = t\}$ such that its first due payments drop under AUD should coincide with the start of the new reform.[16] Its inflow is six months before the start of PARE.[17] The choice of the older cohort (the untreated) is more flexible as we do not need to observe the compliance. The main restrictions have macroeconomic considerations: a good choice of a cohort does not violate the randomization assumption. Busyness cycles or mass layoffs due to bankruptcies of big firms are examples for possible causes for structural changes in the distribution of heterogeneity in the unemployment inflow over time. Bearing these considerations in mind, we choose the cohort of the untreated to be 3 months older.[18] This time lag is long enough for policy analysis, as typically a big part of the exits occurs in the first 3 months. Moreover, the cohorts begin their unemployment spells in a fairly economically stable time interval, see subsection 1.4.4 for a discussion.
With these choices we end up with 537 (311) spells in the treated (nontreated) cohort. From these 116 (76) are censored. In the younger cohort there are 250 compliers (the compliance

[16] One can take also a subsequent digression period, but at the cost of having less observations.

[17] The time length from inflow until the first digression day can vary depending on characteristics of the unemployed, such as number of working days in the last twelve months, age, etc., see Freyssinet (2002) for details.

[18] This has an implication for the time interval of comparison. Conditional on survival up to 6 months, one can compare the two cohorts only in an interval of 3 months. After the 3rd month, the older cohort will also receive the treatment, and one would no longer compare treated with untreated.

indicator for the second cohort is not of interest).

1.4.3 Estimation results

We now turn to our main results. With the choice of cohorts described in 1.4.2 the treatment effect which we estimate is equal to

$$TE(6,a) = P(T(6) \in [6, 6+a) \mid T(6) \geq 6) - P(T(9) \in [6, 6+a) \mid T(9) \geq 6), \tag{1.4.1}$$

where a varies between 0 and 3 months. The upper limit three months follows from the time difference of the inflows of the two cohorts. Any comparison beyond this interval would involve two treated groups. The treatment effect gives thus the difference in the probabilities to find a job in the interval $[6, 6+a)$, conditionally on surviving up to the 6th month, between the old and the new system. Letting a go to 0 would give a comparison of hazards, which is not very informative as it consists of a single point.

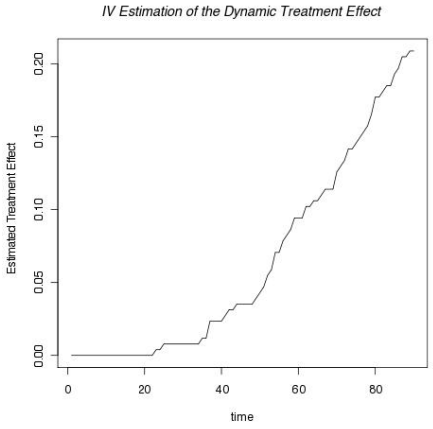
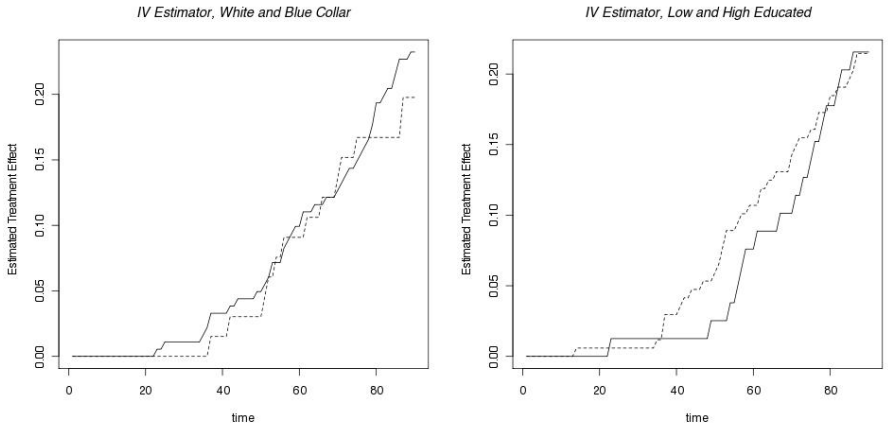


Figure 1.1: An IV estimator of the treatment effect. Time measured in days.

The result is shown on figure 1.1. On the x-axis time is measured in days. Each (x, y)-point represents a pair $(a, ATE(a))$. The estimated treatment effect is positive and increasing which indicates that the program was effective. The estimates are a. e. significant. The results are similar for different subpopulations, see figures 1.2a and 1.2b. Our results are compatible with the findings in the existing literature. Crépon, Ferracci, and Fougère (2007) find that training does not accelerate exit out of unemployment but increases the length of the subsequent employment spell. The authors' finding is related to the idea that training increases human capital and improves matching process between firms and unemployed. Crépon, Dejemeppe, and Gurgand (2005) find that three out four counselling schemes have a positive effect on the unemployment hazard. One possibility is that jointly these two types of services offset the negative impact of the generous benefit system. It must be noted that our evaluation subsumes several different treatments into one single treatment. Not all of the compliers received the full treatment. A big percentage exit unemployment after the first meeting with the caseworker. We interpret our evaluation as averaging over the different treatment schemes. Furthermore, even if there was no real monitoring, it is likely

Figure 1.2: Estimates conditional on qualification and education

(a) White vs Blue collar (dashed) (b) Low vs High educated (dashed)



that the regular meetings with the caseworker were perceived as monitoring. Averaging over different treatment types is not uncommon in the literature, see for example Blundell, Dias, Meghir, and Reenen (2004) and Van den Berg, Bozio, and Dias (2014) for evaluations of the New Deal for Young People program in the UK.

1.4.4 The validity of the assumptions

We start with the randomization assumption A3. To verify that the cohorts are similar at their inflows, we compare

1. the distributions of the observed characteristics,
2. the layoff reasons and
3. the macroeconomic conditions

of the two cohorts at the points in time of their inflow. First, we perform chi-square test for equality of distributions of level of education, years of experience, number of children and gender. The corresponding p-values are 0.6037, 0.98, 0.5112 and 0.581, which indicates that the differences between these distributions are statistically insignificant. This is reflected in their histograms, see figures 1.3a, 1.4a, 1.5a, 1.6a. Second, the same test is performed also for the layoff reasons. The null (equality of distributions) is rejected, but in this case this could be due to the large number of categories and small number of observations in each category. A histogram of aggregated categories indicates that the cohorts are indeed similar, see figure 1.7a. Third, the next we show the average level of unemployment in the administrative region Îll de France in the first three quarters of 2001 is constant and equal to 6.4%, which is evidence for a fairly stable macroeconomic environment[19].

[19]Source: http://www.insee.fr/en/bases-de-donnees/bsweb

Figure 1.3: Histograms level of education

(a) Treated (b) Untreated

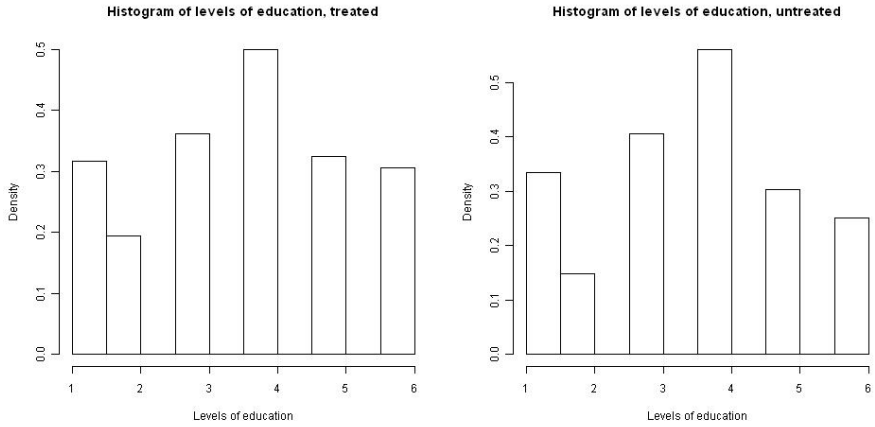
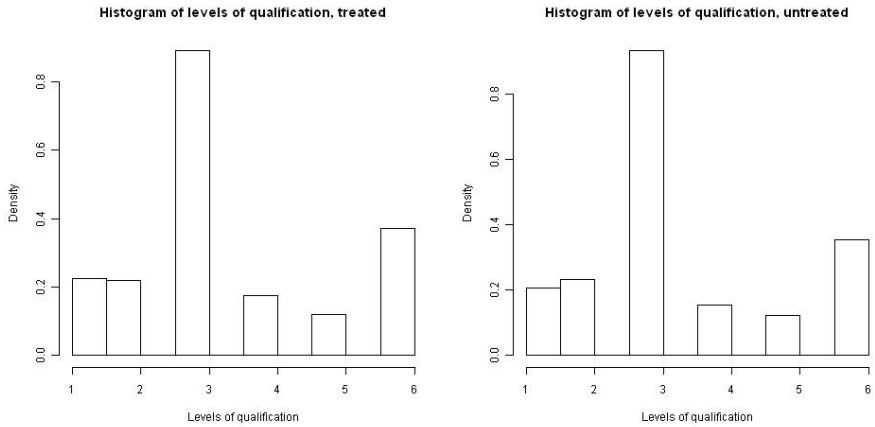


Figure 1.4: Histograms level of qualification

(a) Treated (b) Untreated



Next, the "no anticipation" assumption is fulfilled when individuals do not anticipate the moment in time of treatment or do not act upon this information, see for a discussion Abbring and van den Berg (2003). Although it was known that a reform is going to take place, there was a lot of debate and uncertainty over its content. Unemployed were informed about the exact content and launch date on the 18th of June 2001, that is, less than two weeks before the start of the program, so they had practically no time to react upon this information, see Freyssinet (2002). Further, when an individual decides to switch to the new system, the assignment to a specific treatment depends mostly on the social worker in charge and on the slots available, so that the unemployed has no knowledge of it in advance, see also Crépon,

Figure 1.5: Histograms of years of experience

(a) Treated (b) Untreated

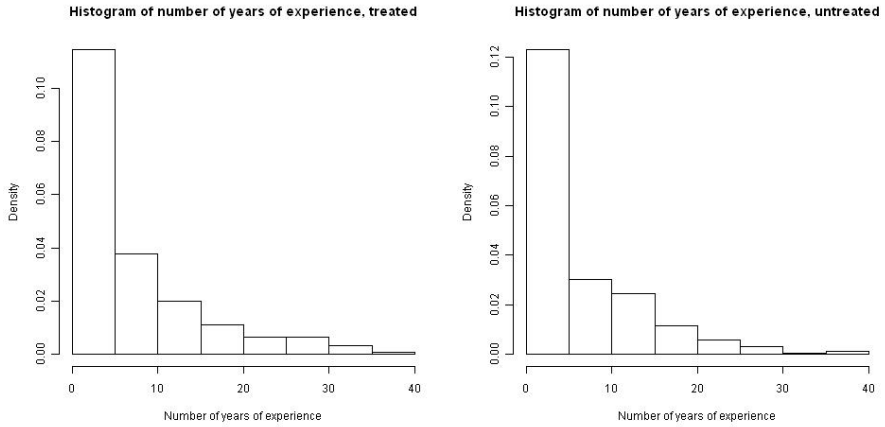
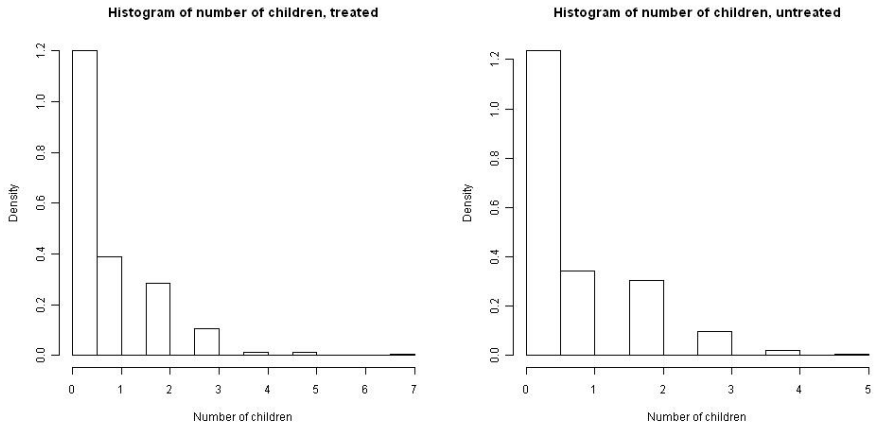


Figure 1.6: Histograms of number of children

(a) Treated (b) Untreated

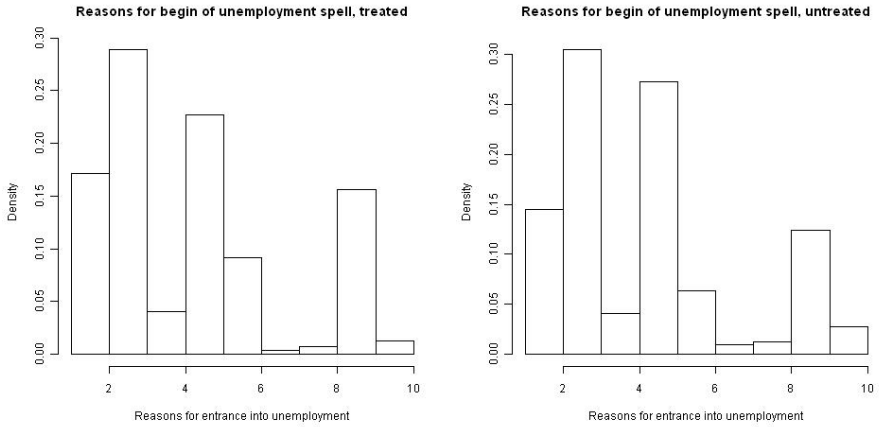


Dejemeppe, and Gurgand (2005). Combined with a very short time span between assignment and launch of a treatment is very short, which precludes acting upon the anticipation.

The last important assumption is that of independent censoring. It cannot be tested directly, as revealed by a nonidentification result of Tsiatis (1975). Over 70% of all censored spells are contributed by the censoring categories ”no control”, ”other cases” and ”other stop of search”. Since there is no further information for these cases, it appears plausible to assume independence. In subsection 1.4.6 we conduct a simulation study, in which plausible deviations from the independence assumptions are generated. It turns out that the estimator is robust towards such violations.

Figure 1.7: Histograms of layoff reasons

(a) Treated (b) Untreated



Remark: one implicit implication of the assumptions A2 and A4 is that noncompliers who refuse the treatment behave in the same way as if they were not assigned to the treatment. It is plausible particularly in the cases, in which individual do not comply **in order** not to change their behavior. In the PARE reform, it is plausible to assume that individuals do not comply because they anticipate a soon exit and because they want to avoid higher search effort or other related participation costs, see the next subsection for an analysis of noncompliance. In both cases, assignment to the treatment together with a selection out of it is not likely to change their behavior.

As a summary, we can conclude that the assumptions adopted for identification and estimation of the treatment effect of the PARE reform can be considered as plausible.

1.4.5 Analysis of endogeneity

In this subsection we tackle the static endogeneity issue arising from noncompliance. Noncompliance is important not only for the evaluation of a program but also in the light of its effectiveness. The non take-up of a policy often reduces the effectiveness of a program, see e. g. Blasco (2009). It is therefore important to understand what drives noncompliance.

We start with an estimation of the naive treatment effect 1.3.11. The corresponding estimator is defined as

$$\widehat{NE}(t,a) := \widehat{P}(T \in [t, t+a) \mid S = t, Z = t) - \widehat{P}(T \in [t, t+a) \mid S = \infty, Z = t), \tag{1.4.2}$$

where t is equal to 6 months and a varies between 1 day and 3 months and the separate probabilities are estimated with a Kaplan-Meier estimator. 1.4.2 amounts to a direct comparison of the average outcome of compliers and noncompliers from the cohort $\{Z = 6\}$. The estimate is shown in figure 1.8. It is positive and increasing until the 80th day after treatment (which is the 260th day of unemployment), and then slightly decreasing. At the first 40 days after treatment the effect is practically zero, at its maximum it is around 0.08,

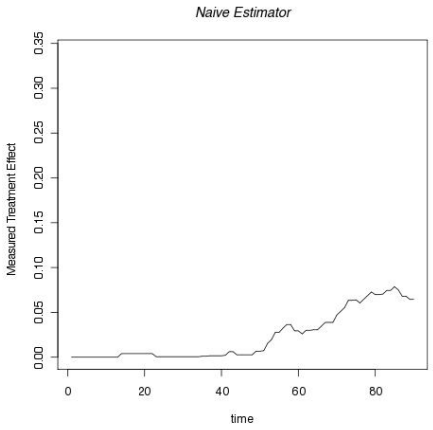


Figure 1.8: A naive estimator: noncompliers as control group. Time measured in days.

and at day 60 after treatment (that is, after 8 months of unemployment) around 0.025. This implies that the probability for a complier to find a job before the end of the first month after treatment, conditional on having been unemployed for 6 months, is almost the same as for a noncomplier, before the end of the second month it is with 2.5 percentage points higher, and at its peak it is 8 pp. higher.[20] As a result, if we evaluate PARE using the naive estimator 1.4.2, we would conclude that the reform was beneficial for the duration of unemployment but that the effect is rather modest.

Using the methods developed in section 1.3.5, we answer now the following questions:

- Is the non take-up of PARE driven by an endogenous selection?
- If yes, in which direction is the bias of the naive estimator caused by this endogenous selection?

To answer the first question, we perform the simplified test for exogeneity from section 1.3.5. The null hypothesis is

$$H_0 : \qquad \bar{F}_1(6 \mid S = \infty) - \bar{F}_2(6). \tag{1.4.3}$$

It amounts to comparing the survival function at $t = 6$ of the noncompliers from cohort $\{Z = 6\}$ with the survival function at $t = 6$ of the whole cohort $\{Z = 9\}$. The test statistics is defined as $\mathscr{T} = \widehat{\bar{F}}_1(T > 6 \mid S = \infty) - \widehat{\bar{F}}_2(6)$, where $\widehat{\bar{F}}_i, i = 1, 2$ are the Kaplan-Meier estimators of $\bar{F}_i$.

The test rejects the null at 5 % level. As a result, untreated compliers and noncompliers are significantly different in terms of potential outcomes, which induces a static selection bias in the naive estimator. To evaluate the bias of 1.4.2, we plot $\widehat{TE}(6, a)$ and $\widehat{NE}(6, a)$ for a varying between 0 and 3 months. The result is shown in figure 1.9. $\widehat{NE}(6, a)$ has a negative bias for all a. To find the reason for this negative bias, it is helpful to interpret it in the frame

[20] Controlling for observed covariates such as gender, education and type of job (white vs. blue collar) yields similar results, see figures 1.14, 1.15a and 1.15b in appendix 1.6.3. Due to the small sample size available, we conduct the study mainly unconditionally.

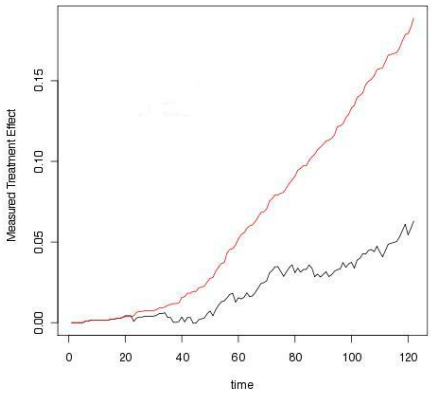


Figure 1.9: Comparison of the IV estimator (red line) and the naive estimator (black line). Time measured in days

of existing studies on policy take-up, see e.g. Moffit (1983), Currie (2004) and Blasco (2009). An empirical analysis of the take-up of the PARE reform is done by Blasco (2009), who also uses a theoretical model. She finds stigma, informational issues and the expectation of a soon exit to be the main reasons for noncompliance. One explanation for the negative bias in line with Blasco (2009) would be therefore the expectation of a short spell among noncompliers. Individuals who anticipate to find quickly a job or who have even already signed a contract at the time of the reform start would be reluctant to comply since they wouldn't benefit from the generosity of the new program. Thus, there is a selection of quick exits into the group of noncompliers which leads to the negative bias of the naive estimator.

1.4.6 Dependent Censoring: a Simulation Study

To assess the impact of the assumption of independent censoring, a small simulation study is conducted. Deviations from $C \perp S$ and $C \perp T$ are constructed, where C again is a censoring random variable. The first one influences the estimator of the probability to be a complier,

$$P(S = t \mid T \geq t, X, Z = t),$$

while the second one influences the estimator of the difference

$$P(T \in [t, t+a) \mid T \geq t, X, Z = t) - P(T \in [t, t+a) \mid T \geq t, X, Z = t').$$

We are interested in their marginal impacts as well as in the influence of their interplay. Two cohorts are simulated, the treated and the nontreated, each with 10000 individuals. Both cohorts consist of compliers and noncompliers and in each cohort the probability to be a complier is 80%. Noncompliers dominate stochastically the compliers when both groups have not received the treatment. This reflects our finding in section 1.4.5 that noncompliance might occur due to the expectation of a short spell. The treatment is obtained by the compliers of the first cohort on the 20th day after inflow and it shifts their duration distribution from $N(60, 15)$

to $N(30,10)$ in line with the estimation results from section 1.4.3[21]. The noncompliers are not influenced by the treatment and have a duration distribution $N(45,15)$. The compliers from the second cohort do not receive the treatment too. Their duration distribution is equal to the duration distribution of the compliers of cohort 1 before treatment, $N(60,15)$. Figure

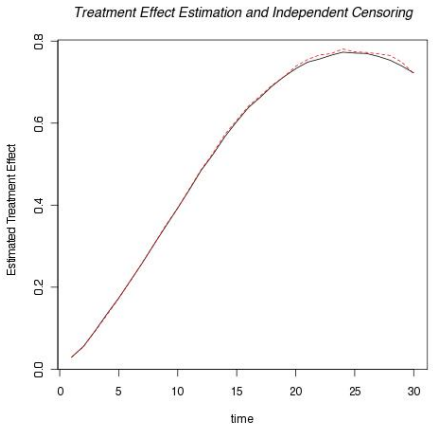


Figure 1.10: An IV estimator of the treatment effect. Time measured in days. Day 0 corresponds to the day of treatment (day 20).

1.10 shows the theoretical treatment effect, depicted by the thick black line. The dashed red line represents the IV estimator in a case with independent censoring with a distribution $N(40,10)$ (the second argument is henceforth the standard deviation) . This is the benchmark estimator.
Next, a dependence of the censoring on the compliance is introduced. The different choices of distributions are described in table 1.1.

Table 1.1: Simulation of dependences between censoring and compliance

Line description	Censoring distribution compliers	Censoring distribution noncompliers
Green dashed line	N(30,15)	N(50,15)
Red dotted line	N(30,15)	N(40,15)
Blue long dashed line	N(40,15)	N(30,15)
Grey two dashed line	N(50,15)	N(30,15)

Notes: The second argument of the normal distribution is its standard deviation

The resulting estimators are shown in figure 1.11. The solid black line is theoretical effect.

[21]Negative values are replaced by their absolute values.

The figure reveals the relationship between bias of the treatment effect and dependence of

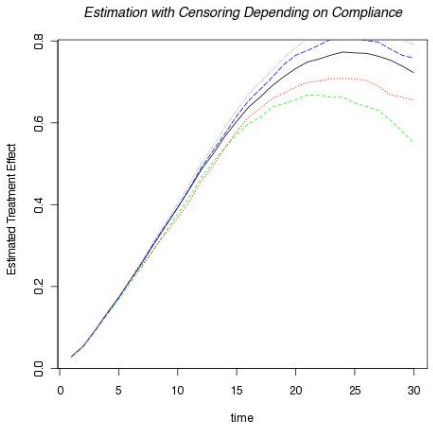


Figure 1.11: An IV estimator of the treatment effect. Time measured in days. Day 0 corresponds to the day of treatment (day 20). The black solid line is the theoretical treatment effect. Different curves correspond to different dependences of censoring and compliance, see table 1.1. The solid black line is theoretical effect.

censoring and compliance. When the compliers are at higher risk of censoring, the treatment effect is (a. e.) underestimated. The higher this discrepancy in the risk exposure, the bigger the bias. Similarly, when the noncompliers are at higher risk of censoring, the treatment effect is overestimated.

Next, the relationship between bias and time dependence of the censoring is exploited. We simulate three different levels of dependence. In all three cases long spells have a higher risk of being censored than short spells. This is in line with typical situations in applied survival analysis. For example, long term unemployed might have smaller incentives to meet criteria (e. g. administrative control of search, regular visits at the agency, etc.) to stay on an unemployment insurance list. The three specifications are defined in table 1.2. Each row represents one specification.

The corresponding estimators are depicted in figure 1.12. Approximately until day 15 the IV estimator performs fairly well in all three cases. Afterwards it underestimates the treatment effect. The bias increases in absolute value with increasing time dependence (defined as the difference in the means in the two groups of spells).

It is interesting to simulate and analyze a combination of these two types dependences. We simulate four patterns of such an interplay. The concrete distributions are described in table 1.3. The results are shown in figure 1.13. The blue and the grey lines are closer to the theoretical effect than the other two estimators. This indicates, that a violation in the censoring assumption $C \perp S$ might partially offset a violation in the assumption $C \perp T$. This is a novel result.

In the French labor market reform it is difficult to argue which type of dependence there

Table 1.2: Simulation of dependences between censoring and time

Line description	Censoring distribution $T \leq 40$	Censoring distribution $T > 40$
Green dashed line	N(40,20)	N(30,20)
Red dotted line	N(40,20)	N(25,20)
Blue long dashed line	N(40,20)	N(20,20)

Notes: The second argument of the normal distribution is its standard deviation

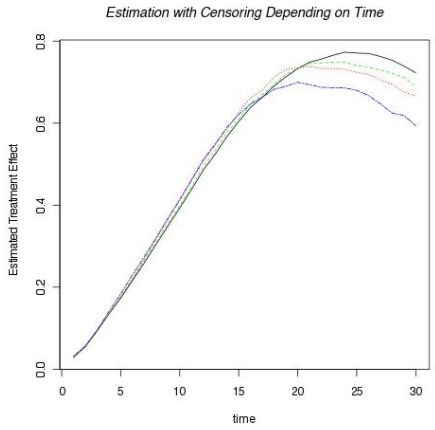


Figure 1.12: An IV estimator of the treatment effect. Time measured in days. Day 0 corresponds to the day of treatment (day 20).The black solid line is the theoretical treatment effect. Different curves correspond to different dependences of censoring and time, see table 1.2. The solid black line is theoretical effect.

is likely to be. Noncompliers contain many quick exits, and if longer spells have a higher censoring risk than shorter spells, than noncompliers should be less exposed to censoring than compliers. This would correspond to the fourth case of table 1.3. Thus the simulation results provide evidence, that the IV estimator is robust to a violation in the independent censoring assumption.

Table 1.3: Simulation of dependences between censoring and compliance and time

Line description	K, $T \leq 30$	K, $T > 30$	N, $T \leq 30$	N, $T \leq 30$
Green dashed line	N(50,20)	N(30,20)	N(30,20)	N(20,20)
Red dotted line	N(40,20)	N(30,20)	N(30,20)	N(20,20)
Blue two dashed line	N(30,20)	N(20,20)	N(40,20)	N(30,20)
Grey long dashed line	N(30,20)	N(20,20)	N(50,20)	N(30,20)

Notes: K stays for compliers, N for noncompliers.

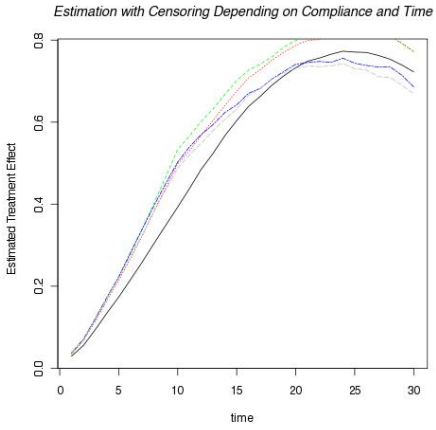


Figure 1.13: An IV estimator of the treatment effect. Time measured in days. Day 0 corresponds to the day of treatment (day 20). The black solid line is the theoretical effect in the absence of censoring. Different curves correspond to different dependences of censoring and time, see table 1.2. The solid black line is theoretical effect.

1.5 Summary and Discussion

In this paper we developed a nonparametric IV framework for the evaluation of dynamic treatment effects. Our methods solve the problems of dynamic and static endogeneity and allow for censoring. The corresponding estimators have a natural interpretation and are related to the Wald-type statistics. We also suggest a framework for analysis of noncompliance. We used our methods to evaluate the French labor market reform PARE. The estimated effect of the reform on the conditional survival function of the unemployment variable is positive, which is in line with the findings in the existing literature. In an exhaustive study, we showed that the assumptions for our approach are valid. Our results reveal that neglecting of endogeneity would lead to a negative bias. An interesting question for future research

would be to incorporate equilibrium effects. Comprehensive policy reforms are likely to induce equilibrium effects through positive or negative externalities. It is often desirable to distinguish between the direct effects of a reform and the equilibrium effects. More work on this topic has to be done.

1.6 Appendix

1.6.1 Proofs of propositions

Proof of Proposition 1.3.1

1. First we show that from the no anticipation assumption the following result holds:

$$P(T(t) \geq t \mid X, S(t) = t) = P(T(t') \geq t \mid X, S(t) = t). \tag{1.6.1}$$

This is so because

$$\overset{\text{No anticipation}}{=} \begin{array}{l} P(T(t) \geq t \mid X, S(t) = t, V) = \exp(-\Theta_{T(t)}(t \mid X, S(t) = t, V)) = \\ \exp(-\Theta_{T(t')}(t \mid X, S(t) = t, V)) = P(T(t') \geq t \mid X, S(t) = t, V) \end{array}$$

so that we obtain

$$\begin{aligned} & P(T(t) \geq t \mid X, S(t) = t) = \mathbb{E}\left[I_{\{T(t)\geq t\}} \mid X, S(t) = t\right] = \\ = \; & \mathbb{E}\left[\mathbb{E}\left[I_{\{T(t)\geq t\}} \mid X, S(t) = t, V\right] \mid X, S(t) = t\right] = \\ = \; & \mathbb{E}\left[P(T(t) \geq t \mid X, S(t) = t, V) \mid X, S(t) = t\right] = \\ = \; & \mathbb{E}\left[P(T(t') \geq t \mid X, S(t) = t, V) \mid X, S(t) = t\right] = \\ = \; & \mathbb{E}\left[\mathbb{E}\left[I_{\{T(t')\geq t\}} \mid X, S(t) = t, V\right] \mid X, S(t) = t\right] = P(T(t') \geq t \mid X, S(t) = t) \end{aligned}$$

where $I_{\{T(s)\in B\}}$ is an indicator function equal to 1 when $T(s) \in B$ (of course from these steps we also see that $P(T(t) \geq t \mid X, S(t) = t, V) = P(T(t') \geq t \mid X, S(t) = t, V)$).

2. Next, using result 1.6.1, we show $F_{V|T(t)\geq t,X,S(t)=t} = F_{V|T(t')\geq t,X,S(t)=t}$. Let B be a Borel set. With result 1.6.1, it holds

$$P(V \in B \mid T(t') \geq t, X, S(t) = t) = P(V \in B \mid T(t) \geq t, X, S(t) = t).$$

3. Now we show $F_{V|T(t)\geq t,X,S(t)=t} = F_{V|T\geq t,X,S=t,Z=t}$. First we observe that $Z \perp\!\!\!\perp \{T(s), S(z)\} \mid X, V$ and $Z \perp\!\!\!\perp V \mid X$ together imply $Z \perp\!\!\!\perp \{T(s), S(z)\} \mid X$ (Weak Union, see Pearl (2000)). Then, we have

$$P(V \in B \mid T(t) \geq t, X, S(t) = t) = \frac{P(V \in B \mid X, S(t) = t)P(T(t) \geq t \mid X, S(t) = t, V \in B)}{P(T(t) \geq t \mid X, S(t) = t)}.$$

We study the separate components of the right-hand side of the last expression.

(a) With assumptions A3 and A4, it holds

$$P(V \in B \mid X, S(t) = t) = P(V \in B \mid X, S = t, Z = t).$$

(b) Further,

$$P(T(t) \geq t \mid X, S(t) = t, V \in B) = P(T \geq t \mid X, S = t, V \in B, Z = t).$$

(c) Using $Z \perp\!\!\!\perp \{T(s), S(z)\} \mid X$ instead of $Z \perp\!\!\!\perp \{T(s), S(z)\} \mid X, V$, we obtain

$$P(T(t) \geq t \mid X, S(t) = t) = P(T \geq t \mid X, S = t, Z = t)$$

So finally we get the equality

$$\begin{aligned} & P(V \in B \mid T(t) \geq t, X, S(t) = t) = \\ = \quad & \frac{P(V \in B \mid X, S = t, Z = t) P(T \geq t \mid X, S = t, V \in B, Z = t)}{P(T \geq t \mid X, S = t, Z = t)} = \\ = \quad & P(V \in B \mid T \geq t, X, S = t, Z = t) \end{aligned}$$

□

Proof of corollary 1.3.2.1
With proposition 1.3.1,

$$\begin{aligned} & TE(a,t) = \mathbb{E}\Big[P(T(t) \in [t, t+a) \mid T(t) \geq t, X, V, S(t) = t) \mid T(t) \geq t, X, S(t) = t\Big] - \\ - \quad & \mathbb{E}\Big[P(T(t') \in [t, t+a) \mid T(t') \geq t, X, V, S(t) = t) \mid T(t') \geq t, X, S(t) = t\Big] = \\ = \quad & P(T(t) \in [t, t+a) \mid T(t) \geq t, X, S(t) = t) - P(T(t') \in [t, t+a) \mid T(t') \geq t, X, S(t) = t). \end{aligned}$$

Proof of proposition 1.3.2
First, consider the conditional distribution of the duration variable for the treatment group, $F_{T(t)|T(t)\geq t, X, S(t)=t}$. Set $B = [t, t+a)$. With randomization and consistency, it holds

$$\begin{aligned} P(T(t) \in B \mid X, S(t) = t) &= P(T \in B \mid X, S = t, Z = t) \\ P(T(t) \geq t \mid X, S(t) = t) &= P(T \geq t \mid X, S = t, Z = t), \end{aligned}$$

so that

$$P(T(t) \in B \mid T(t) \geq t, X, S(t) = t) = P(T \in B \mid T \geq t, X, S = t, Z = t) \tag{1.6.2}$$

where the r.h.s of 1.6.2 consists only of observables. In the simplified notation of section 1.3.2, this gives us $F_{C,1}$, that is, the average outcome of the treated t-compliers. To obtain the average outcome of the nontreated t-compliers $F_{C,0}$, write

$$\begin{aligned} & P(T \in B \mid T \geq t, X, Z = t') \qquad (1.6.3) \\ = \quad & P(T \in B \mid T \geq t, X, Z = t', S(t) = t) P(S(t) = t \mid T \geq t, X, Z = t') + \\ + \quad & P(T \in B \mid T \geq t, X, Z = t', S(t) = \infty) P(S(t) = \infty \mid T \geq t, X, Z = t'), \end{aligned}$$

where $P(T \in B \mid T \geq t, X, Z = t')$ contains only observables. Our identification proof contains the following steps:

1. Show that $P(T \in B \mid T \geq t, X, Z = t', S(t) = t)$ is equal to $P(T(t') \in B \mid T(t') \geq t, X, S(t) = t)$ and is therefore the expression we want to identify.

2. Show that

$$P(T \in B \mid T \geq t, X, Z = t', S(t) = \infty) = P(T \in B \mid T \geq t, X, Z = t, S(t) = \infty),$$

   i.e. the noncompliers of the two cohorts have identical potential duration distributions at t.

3. Show that $P(S(t) = t \mid T \geq t, X, Z = t')$ and $P(S(t) = \infty \mid T \geq t, X, Z = t')$ are identified (these are the proportions of compliers and noncompliers at t).

4. Solve for $P(T \in B \mid T \geq t, X, Z = t', S(t) = t)$ in 1.6.3 and relate it to osbervables via steps 1-3.

Let's proof these points. First, It holds

$$\begin{aligned}
&P(T \in B \mid X, Z = t', S(t) = t) = \\
&= P(T \in B \mid X, Z = t', S(t) = t, S = t')P(S = t' \mid X, Z = t', S(t) = t) + \\
&+P(T \in B \mid X, Z = t', S(t) = t, S = \infty)P(S = \infty \mid X, Z = t', S(t) = t) = \\
&\overset{\text{Consistency}}{=} P(T(t') \in B \mid X, Z = t', S(t) = t, S = t')P(S = t' \mid X, Z = t', S(t) = t) + \\
&+P(T(\infty) \in B \mid X, Z = t', S(t) = t, S = \infty)P(S = \infty \mid X, Z = t', S(t) = t) = \\
&\overset{\text{Consistency}}{=} P(T(t') \in B \mid X, Z = t', S(t) = t, S(t') = t')P(S(t') = t' \mid X, Z = t', S(t) = t) + \\
&+P(T(\infty) \in B \mid X, Z = t', S(t) = t, S(t') = \infty)P(S(t') = \infty \mid X, Z = t', S(t) = t) = \\
&\overset{\text{No anticipation}}{=} P(T(t') \in B \mid X, Z = t', S(t) = t, S(t') = t')P(S(t') = t' \mid X, Z = t', S(t) = t) + \\
&+P(T(t') \in B \mid X, Z = t', S(t) = t, S(t') = \infty)P(S(t') = \infty \mid X, Z = t', S(t) = t) = \\
&= P(T(t') \in B \mid X, Z = t', S(t) = t) \overset{\text{Randomization}}{=} P(T(t') \in B \mid X, S(t) = t).
\end{aligned}$$

Next, using exactly the same steps as in the previous point, it follows

$$\begin{aligned}
&P(T \in B \mid T \geq t, X, Z = t', S(t) = \infty) = P(T \in B \mid T \geq t, X, Z = t, S(t) = \infty) \qquad (1.6.4) \\
= \;& P(T \in B \mid T \geq t, X, Z = t, S = \infty).
\end{aligned}$$

Further, with assumptions A2-A4

$$P(T \geq t \mid X, Z = t') = P(T \geq t \mid X, Z = t) \qquad (1.6.5)$$

for all $t' \geq t$.[22] Its validity can be proved as follows:

$$
\begin{aligned}
& P(T \geq t \mid X, Z = t') = P(T \geq t \mid X, Z = t', S = t')P(S = t' \mid X, Z = t') + \\
+ \quad & P(T \geq t \mid X, Z = t', S = \infty)P(S = \infty \mid X, Z = t') = \\
\overset{\text{Consistency}}{=} \quad & P(T(t') \geq t \mid X, Z = t', S(t') = t')P(S(t') = t' \mid X, Z = t') + \\
+ \quad & P(T(\infty) \geq t \mid X, Z = t', S(t') = \infty)P(S(t') = \infty \mid X, Z = t') = \\
\overset{\text{Randomization}}{=} \quad & P(T(t') \geq t \mid X, S(t') = t')P(S(t') = t' \mid X) + \\
+ \quad & P(T(\infty) \geq t \mid X, S(t') = \infty)P(S(t') = \infty \mid X,) = \\
\overset{\text{No anticipation}}{=} \quad & P(T(\infty) \geq t \mid X, S(t') = t')P(S(t') = t' \mid X) + \\
+ \quad & P(T(\infty) \geq t \mid X, S(t') = \infty)P(S(t') = \infty \mid X,) = \\
= \quad & P(T(\infty) \geq t \mid X)
\end{aligned}
$$

If we set $t' = t$ and follow exactly the same lines we get

$$P(T \geq t \mid X, Z = t) = P(T(\infty) \geq t \mid X)$$

and finally $P(T \geq t \mid X, Z = t') = P(T \geq t \mid X, Z = t)$.

Using relation 1.6.5 together with 1.6.4 and randomization, we obtain:

$$P(S(t) = \infty \mid T \geq t, X, Z = t') = P(S(t) = \infty \mid T \geq t, X, Z = t).$$

Taking into account that

$$P(S(t) = t \mid T \geq t, X, Z = t') = 1 - P(S(t) = \infty \mid T \geq t, X, Z = t')$$

we finally obtain the equalities

$$P(S(t) = \infty \mid T \geq t, X, Z = t') = P(S = \infty \mid T \geq t, X, Z = t), \tag{1.6.6}$$

$$P(S(t) = t \mid T \geq t, X, Z = t') = P(S = t \mid T \geq t, X, Z = t). \tag{1.6.7}$$

Taking into account the results 1-3 and solving 1.6.3 for $F_{C,0}$, we obtain

$$
\begin{aligned}
& P(T(t') \in B \mid T(t') \geq t, X, S(t) = t) = \\
& \frac{P(T \in B \mid T \geq t, X, Z = t') - P(T \in B \mid T \geq t, X, Z = t, S = \infty)P(S = \infty \mid T \geq t, X, Z = t)}{P(S = t \mid T \geq t, X, Z = t)}.
\end{aligned}
$$

Finally, the treatment effect is equal to $F_{C,1} - F_{C,0}$ which after simplification is equal to

$$\frac{P(T \in B \mid T \geq t, X, Z = t) - P(T \in B \mid T \geq t, X, Z = t')}{P(S = t \mid T \geq t, X, Z = t)}.$$

□

Proof of proposition 1.3.5

[22] This we will refer to this result as *empirical no anticipation* relation.

For notational simplicity we drop the dependence on 0 and x_0. First note, that the results of Theorem 1 Nielsen and Linton (1995) remain valid at the boundary when we replace the symmetric kernel k with its boundary counterpart k_+ and adapt the constants. The validity of 1.3.5 i) follows from $\sqrt{nb^{q+1}}((\widehat{\Psi} - \Psi^*) = \frac{\sqrt{nb^{q+1}}}{\widehat{p_1}}((\widehat{\theta_1} - \theta_1^*) - (\widehat{\theta_2} - \theta_2^*))$, the independence of $(\widehat{\theta_1} - \theta_1^*)$ and $(\widehat{\theta_2} - \theta_2^*)$, and the adapted proof of Theorem 1 i) in Nielsen and Linton (1995). Next, it holds

$$b^{-2}(\Psi^* - \Psi) = \frac{b^{-2}}{\widehat{p_1}}((\theta_1^* - \theta_1) - (\theta_2^* - \theta_2)) + b^{-2}(\theta_1 - \theta_2)(\frac{1}{\widehat{p_1}} - \frac{1}{p_1}). \tag{1.6.8}$$

The second term on the right-hand side of 1.6.8 is equal to $o_p(1)$ when b is of order $O(n^{-1/(q+5)})$ or $o(n^{-1/(q+5)})$. Result 1.3.5 ii) follows with Theorem 1 b) in Nielsen and Linton (1995). Finally, 1.3.5 iii) follows directly from the adapted proof of Theorem 1 c) Nielsen and Linton (1995) and the continuous mapping theorem.

1.6.2 Description of variables

The variables used in our application have been constructed in the following way:

- The variable **age** gives the age at the begin of the unemployment spell and is defined as the year in which the spells begins minus the year of birth.
- **Marital status** consists of four categories: single, married, divorced and widowed.
- the variable for **educational level** summarizes the 31 categories used in the administrative data set into 6 categories according to the highest degree attained. The correspondence is roughly as follows: value 1 if the degree is in niveau I and II (university degree, maîtrise and licence), value 2 if the degree is in niveau III - BTS and DUT (brevet de technicier supérieur and diplôme univeritaire de technologie, respectively, both technical degrees obtained in 2 years after high school), value 3 for all Baccalauréat (high school degree, the general part of lycée) diplomas and for all dropouts from niveau III, 4 for all BEP ,CEP (professional Baccalauréat, specialised part of lycée) and all dropouts from Baccalauréat, 5 for BEPC (brevet d'études du premier cycle, junior high school), and 6 for below.
- The variable **experience** states the number of years of experience in the job (type and position), which the individual is looking for. The types of jobs are specified in an administrative nomenclature table (ROME table). There are several hundred different types.
- The **job type** variable contains general information about the type of the activity in the job preceding the current unemployment spell. It summarizes the 9 administrative categories into 6 categories: white collar skilled, white collar unskilled, technical, supervisor (a production team leader) and manager. This summarized categorization is in line with existing literature, see for example Crépon, Ferracci, Jolivet, and Van den Berg (2010). The initial administrative variable is contained in the FH data set. This holds also for the variable, which states which job is the unemployed looking for, while

the following employment type and position is contained in the DADS data set. Unfortunately, there is no clear matching between the variables from the two different data sets, which leads to some unclarity regarding the question whether the unemployed actually found the job he/she was looking for. This restricts our definition of censoring. Therefore, in this application each observation with known job destination is considered uncensored.

- **Censoring indicator**: there are several possibilities, when an observation is considered as censored. These are:
  - when the unemployment spell in the data set is not finished at the time of the data collection, or
  - when the individual exits the labor market. This includes exits to maternity, accident, illness or invalidity, invalidity pension, military service, administrative change of insurance status, attrition because of nonsufficient administrative control, dropout because of nonregular notifications, and other, nonspecified reasons. While reasons such as maternity, military services and invalidity pension are normally known well in advance by the unemployed and can therefore be related to search activity (as well as to compliance behavior), they represent a small fraction of the observations.
- **Unemployment history**: it is constructed as a binary variable which equals 1 if the individual had been already unemployed before the last employment spell. There are various ways to define unemployment history. One example is the total length of previous unemployment spells. Alternatively, one could take the number of unemployment spells, or both. All possibilities suffer from disadvantages. The last possibility seems to provide the most complete information, but it also demands more data, since it provides many different categories. The total length of previous unemployment lacks any information about the lengths of the separate spells, and the number of spells alone doesn't give any information about the length of unemployment. The binary indicator also does not provide any information at all about the dispersion of previous unemployment, but it is easy to understand and requires only two categories, which makes it computationally attractive. Additional, more serious drawback for the other two indicators is, that the data set is left censored: the earliest information about employment is from 1993. This problem is less severe, if one only looks at the indicator of having been unemployed.

1.6.3 Analysis of endogeneity

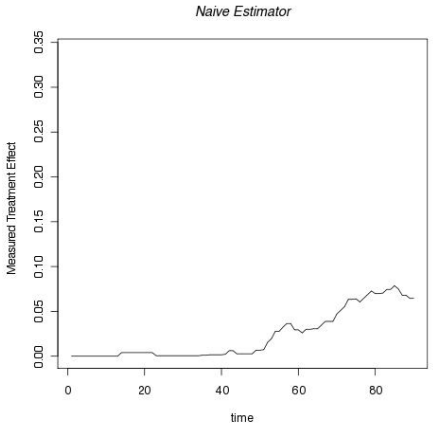
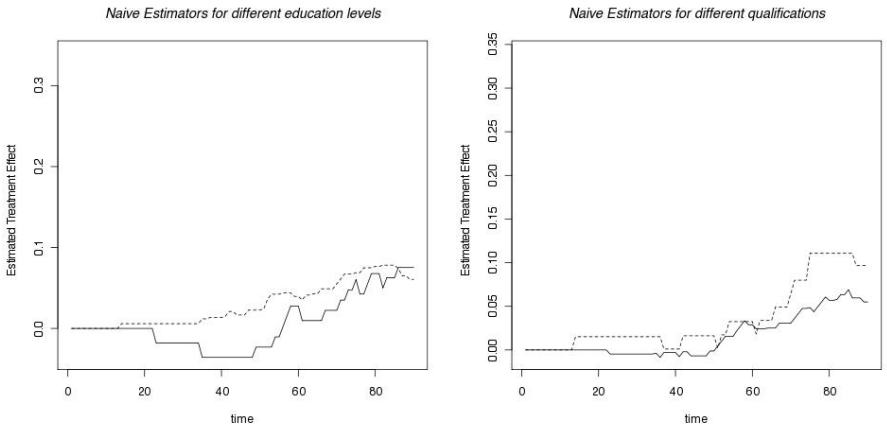


Figure 1.14: A naive estimator. Male vs. Female. Time measured in days.

Figure 1.15: A naive estimator for subgroups

(a) Education. Low educated dashed line (b) Qualification. Blue collars dashed line



Chapter 2

Class size and school performance: a nonparametric IV shape analysis

2.1 Introduction

In this paper, I study the effect of class size on school achievement in two datasets on school achievement: i) a US dataset studied in Cho, Glewwe, and Whitler (2012) applying the random enrollment variation method developed in Hoxby (2000); and ii) an Israeli dataset studied by Angrist and Lavy (1999) with an administrative maximum class-size rule as instrument. These studies represent two influential instrumental variables strategies and are comparable with respect to the age of the evaluated children (3rd - 5th degree) and the type of schools (public schools).

One main difficulty in evaluating the effect of class size on test scores is the endogeneity of class size. Class size might vary with other, typically unobserved, determinants of school success. Examples of such educational inputs are unobserved human capital investments made by the parents, as well as teacher quality. One way in which the endogeneity of the class size has been addressed is by using instrumental variables (IV). However, despite a considerable body of IV literature on this topic, there is still no general consensus on the direction of the effect and its significance, see Averett and McLennan (2004) for an overview.

The main contribution of my paper is to show that model specification error can potentially explain the differences in the literature. My empirical analysis consists of three parts. First, I analyze graphically the regression function using unconstrained and monotonically constrained nonparametric IV estimators. The unconstrained estimator is increasing up to class size of approximately 25 students and then decreasing. The (decreasing) constrained estimator is concave and deviates substantially from the unconstrained counterpart. Second, I develop an empirical test for monotonicity of a regression function under endogeneity and apply it to both data sets. The results of both my graphical and empirical testing analysis indicate that the regression function is non-monotone. Hence, a nonlinear way of modeling the causal relationship is required. Third, I show that second degree polynomials approxi-

mate the causal effect astonishingly well on a range containing 25 % of all observations. To make these graphical findings legal, I test for parametric specifications under endogeneity. Both quadratic and cubic polynomials are shown to be robust specifications.

In a small simulated example I demonstrate how a non-monotone causal relationship together with differences in the class size distributions can potentially explain the controversy that exists within the literature. The magnitude of the model specification error depends sensitively on the range of observed class sizes and the form of the regression function.

Non-monotone class-size effects are difficult to explain with existing theoretical literature. Economic theories generally model class education either as a private good, Brown and Saks (1980), or as a public good subject to congestion, Lazear (2001). In the former case, the teacher pays less attention to each student when the class size increases, whereas in the latter case, the probability of a lesson's impediment (due to a disruption or question) increases with the class size. In both cases, a higher class size leads to a decrease in the quality of the educational process. Theories predicting a positive relationship usually attribute this to interactions with peers, see Schunk (1991) and Sacerdote (2011).[1] The empirical evidence provided in this paper suggests that the overall effect of class size on test scores is a combination of opposite effects that dominate in different class-size ranges. I develop a simple model of an educational production function that can generate the non-monotone pattern found in the empirical analysis.

My paper makes also several contributions to the literature on constrained nonparametric instrumental variable regression.

First, I show that a broad class of constrained penalized minimum distance (PMD) estimators are projections of the corresponding unconstrained estimators on the constrained set with respect to some (weak) norm. This class of estimators includes Tikhonov regularization estimation procedures. The characterization as a projection provides predictions for the properties of constrained estimators. In addition, I develop a framework that reveals a class of constrained Tikhonov estimators as two-step projection (TSP) estimators: in a first step, the dependent variable is projected on the set of unconstrained functions of the independent variable. In a second step, this projection is projected on the set of functions which fulfill the constraint. This characterization provides a theoretical framework for the study of the properties of constrained estimators.

Second, I analyze the properties of constrained PMD estimators based on sieves and kernels. These two frameworks are chosen because they are popular approaches in the nonparametric IV literature, see e.g. Blundell, Chen, and Kristensen (2007) and Chen and Pouzo (2012) for sieves and Darolles, Fan, Florens, and Renault (2011) for kernels. In both cases, I utilize the above discussed projection property to provide sufficient conditions for the consistency of the constrained estimator $\widehat{m}_C$. I show that these conditions are implied by assumptions that are related to both the rate of ill-posedness of the inverse problem and the smoothness of the model solution. Intuitively, a strong instrument and a smooth regression ensure the consistency of $\widehat{m}_C$. Under these assumptions, no further requirements on the constrained set C are necessary (apart from the indirect requirements that $\widehat{m}_C$ exists and that

[1] Lazear (2001) acknowledges that higher class size potentially yields positive peer effects. However, he argues that since increasing class size reduces costs, if students pay the value of the education they receive, adding extra students will generally take place for class sizes where the negative effect of class size dominate the positive one.

the model solution is in C). This is a novel result.[2]

Third, for a series-based estimation procedure, I derive in an exhaustive simulation the optimal cutoff parameter and the optimal regularization constant for different function forms, sample sizes, degree of endogeneity and strength of the instrument. I use my findings to outline a guideline for applied research. The constrained estimator is shown to outperform the unconstrained counterpart, provided that the true regression function is in the constrained set. The advantage increases ceteris paribus both i) with an increasing degree of endogeneity and ii) with a decreasing strength of the instrument. This small-sample weak-instrument simulation result complements the theoretical findings for the behavior of constrained estimators under low endogeneity and/or strong instruments.

Fourth, I show that when the model solution is an inner point of the monotonically constrained set, the unconstrained Tikhonov estimator fullfils the constraint w.p.t.1. As a result, constrained and unconstrained Tikhonov estimators are asymptotically equivalent. Thus, in this case there are no gains from imposing a constraint for the convergence rates.

I utilize the constrained sieves IV estimator in two different ways. First, I use it to perform the graphical analysis of the causal effect of class size on test scores. Second, I use it to construct the empirical ad hoc test for monotonicity. I adapt the framework developed in Breunig (2012) and Breunig (2013) to the case of monotonicity. Intuitively, the test statistic is based on an empirical measure of distance between the data and the "closest" estimated monotone function. Big values of the test statistic indicate deviations from monotonicity. The test is shown in an extensive simulation study to have good consistency and power properties. I construct both alternatives that have a small (non-monotone) dip on a large range and alternatives that have a deep dip on a small interval. In both cases, the procedure achieves the theoretical values when the sample size increases.

The remainder of this paper is structured as follows. In section 2, I relate my paper to the existing literature. In section 3, I discuss the endogeneity of the class size and present the econometric framework. Section 4 develops the econometric theory. I present the simulation results in section 5. In section 6, I conduct my empirical investigation with the Minnesota data set. Section 7 concludes. In the appendix, I prove the assertions from section 4 and analyze the Israeli data set.

2.2 Related Literature

There is a vast and still growing literature on the effect of class size on test scores and IV studies constitute a considerable part of it. Table 2.1 gives a (non-exhaustive) overview of this literature. Three often used instrumental variable strategies are administrative maximum class-size rules, average class size (in each school on a school level or grade level), and random variation in the population. Interestingly, even studies sharing the same instrumental strategy often depart in their findings. For example, Dobbelsteen, Levin, and Oosterbeek (2002) find small positive significant or no significant effects, Hoxby (2000) finds insignificant effects and Angrist and Lavy (1999) and Gary-Bobo and Mahjoub (2006) find negative significant effects, all using a maximum class-size rule. Similarly, Akerhielm (1995) and Bressoux, Kramarz,

[2] typically properties of constrained estimation procedures are developed in the context of closed and convex constraints.

Table 2.1: Overview of empirical IV literature analyzing the effect of class size on test scores

Study	Country data	Instrument
Akerhielm (1995)	U.S. (NELS)	(grade specific) average class size in each school
Angrist and Lavy (1999)	Israel	Administrative maximum class-size rule (Maimonides' rule)
Case and Deaton (1999)	South Africa	District pupil /Teacher ratio, racial composition at district level
Hoxby (2000)	U.S.	Random population variation and maximum class-size rule
Boozer and Rouse (2001)	U.S. (NELS)	Average student/Teacher ratio in the state
Dobbelsteen, Levin, and Oosterbeek (2002)	Netherlands (PRIMA)	Administrative rule on the number of teachers in a school
Bonesrøinning (2003)	Norway	Administrative maximum class-size rule
Gary-Bobo and Mahjoub (2006)	France	Administrative maximum class-size rule
Wößmann and West (2006)	18 countries (TIMMS)	Average class size at respective grade
Bressoux, Kramarz, and Prost (2009)	France	Average class size in the school
Cho, Glewwe, and Whitler (2012)	U.S.	Random population variation

and Prost (2009) find negative significant effects using an average class-size as an instrument, while the study of Wößmann and West (2006) finds in 4 of 18 countries negative, in 1 positive and otherwise insignificant effects using the same instrument. As a further example, Hoxby (2000) finds no significant effects with a random population variation as an instrument, while the estimates in Cho, Glewwe, and Whitler (2012) are negative and significant. All those papers obtain their estimates with a linear specification in a 2SLS framework. I contribute to the IV literature on the effect of class size on test score by showing that a model specification

error might be the source of the discrepancy.

My empirical test utilizes the framework of Breunig (2012) and Breunig (2013) by adapting it to the case of monotonicity. It is also related to the testing procedures in Horowitz (2006) and Horowitz (2012).

This study also builds on the literature on TSP frameworks for constrained estimation. The paper of Mammen, Marron, Turlach, and Wand (2001) develops a TSP framework for simple smooth and the papers of Mammen and Thomas-Agnan (1999) and Mammen, Marron, Turlach, and Wand (2001) characterize constrained smoothing splines as projections. I extend their ideas to the case of Tikhonov regularized estimation procedures.

2.3 The Endogeneity of the Class Size and the Econometric Model

Denote with Y_i the test score of a student i, with X_i the class size of the class of student i and with ϵ_i some other characteristics which are unobservable to the econometrician and which also influence the test scores. An example for ϵ_i is the ability of student i. Consider the model

$$Y_i = m_0(X_i) + \epsilon_i, \tag{2.3.1}$$

where m_0 summarizes the causal effect of X_i on Y_i and is unknown.[3] Alternative definition of this model is with Y_i being the average test score in a class i. This aggregated formulation is often imposed in the literature because of the lack of information on an individual level, see Angrist and Lavy (1999) for a discussion. I will use it throughout this paper without elaborating on the issues related to it. For now suppose that an i.i.d. sample (Y_i, X_i, ϵ_i) is drawn from the distribution of (Y, X, ϵ).

The major difficulty in identifying the causal effect m_0 of class size on test scores is that the class size might be endogenous, that is, $\mathbb{E}[\epsilon \mid X] \neq 0$. The following exposition of the reasons for endogeneity of X is based on the references in section 2.2. The class size is endogenous if there is a systematic difference between the observables along different class sizes. As an example schools with smaller classes might attract better teachers which will lead to a negative estimation bias. In general, the channels through which endogeneity might occur can be divided into two groups: between-schools selection and within-school selection. Between-schools selection occurs for example when

- Schools with overall better resources (such as better teachers) attract more students, which potentially leads to higher class sizes. This would induce ceteris paribus a positive spurious correlation of class size and test scores.
- Parents who invest more into the human capital of their children might also make more efforts to get them into smaller classes (an underlying assumption is that parents actually *believe* that smaller classes are better). This could lead to a negative bias.
- Parents trying to compensate for/reinforce the low/high ability of their child might choose schools with smaller classes.

[3] For expositional simplicity, I abstract in this section from other observed covariates. In the empirical section, I impose and discuss a model that also includes other observed covariates.

- In rural areas schools might have systematically different class sizes than those in urban areas. If also the distributions of the (usually unobserved) ability differ then the estimates would capture these differences.
- In general, well-off families might choose to purchase their residences in areas with better schools. This is referred to as Tiebout sorting. It poses a problem for the identification of m_0 when the better schools systematically differ from the other schools in terms of class size.

Within-school selection can occur for example when

- Teachers/directors assign students with learning difficulties to smaller classes.
- Parents who care more for the school achievement bargain more aggressively with the teachers/school directors to get their children into smaller classes.
- In areas with population from a lower sociological background and lower quality students, teachers/school directors/local policy makers might treat preferentially high skilled students and students from well-off families to prevent them from leaving the area.

The list of possible reasons is not exhaustive. The endogeneity is generated in the interplay of all education stakeholders and in most cases it is difficult to say in which direction the bias would be.

An instrumental variable strategy to tackle the endogeneity issue is to use a variable W which fulfills

$$\mathbb{E}[\epsilon \mid W] = 0. \tag{2.3.2}$$

Combined 2.3.1 and 2.3.2 give the instrumental variable equation

$$\mathbb{E}[Y \mid W] = \mathbb{E}[m_0(X) \mid W]. \tag{2.3.3}$$

Assuming a differentiable regression function, the question whether the class size effect is monotone is equivalent to the question whether $m' \leq 0$ a.e. on the domain of X (or alternatively $m' \leq 0$), where m' denotes the first derivative of m.

The class-size/test-scores regression function is typically modeled in the literature as a linear function. An estimator is obtained, roughly speaking, by inverting the right-hand side and replacing the unknown expectations with their sample counterparts. In most of this paper, I do not impose a parametric form for m_0.

2.4 Shape constraints in minimum distance penalized procedures: estimation and testing

2.4.1 The framework for constrained estimation

Assume model 2.3.1 and assume that there is an observable variable W such that the restriction 2.3.2 holds. Suppose further that $Q := Q(m)$ is a population criterion that is uniquely

minimized at the "true" regression function m_0. An ill-posed inverse problem exists, when there is a sequence of admissible functions (=potential regression functions) $\{m_k\}$ such that $Q(m_k) \to Q(m_0)$ but $d(m_k, m_0) \nrightarrow m_0$, where d is some metric on the space of admissible functions. This is for example the case when the inverse of Q is not continuous (for more general definition of well-posed and ill-posed inverse problems see for example Engl, Hanke, and Neubauer (1996). A typical solution is to regularize the sample criterion with a penalty term. To be more concrete, define the operator T as

$$\begin{aligned} T: \quad & L_X^2 \longrightarrow L_W^2 \\ & \phi \longrightarrow T\phi(w) := \mathbb{E}[\phi(X) \mid W = w], \end{aligned}$$

where $L_X^2 := \{\phi : \mathbb{E}[\phi(X)^2] < \infty\}$ and $L_W^2 := \{g : \mathbb{E}[g(W)^2] < \infty\}$. Suppose that $h(w) := \mathbb{E}[Y \mid W = w]$ is in L_W^2. Then the model 2.3.3 can be written as

$$Tm = h. \tag{2.4.1}$$

In this paper, I consider regularized minimum distance estimators of the type

$$\widehat{m} = \underset{m \in G}{\operatorname{argmin}} \|\widehat{T}m - \widehat{h}\|_H^2 + \alpha_n \|m\|_G^2 = \underset{m \in G}{\operatorname{argmin}} \mathscr{F} m, \tag{2.4.2}$$

where H and G are Hilbert spaces, $\|.\|_G$ is a norm on G and $\|.\|_H^2$ is some minimum distance criterion, for example a norm on H. α_n is a constant that depends on the sample size n (throughout this paper I will omit the index indicating the dependence of the estimator on n) and $\widehat{T}, \widehat{h}$ are estimators of T and h, respectively.

One appealing feature of these estimation procedures is that constraints can be easily imposed on the solution. If C is a subset of G, then the constrained estimator $\widehat{m}_C$ is defined as

$$\widehat{m}_C = \underset{m \in C}{\operatorname{argmin}} \mathscr{F} m. \tag{2.4.3}$$

Whereas the focus of the empirical study is whether $m' \leq 0$ holds, I first consider the broader problem of estimating m_0 under (general) shape constraints in the context of ill-posed inverse problems.

Imposing a shape constraint on an estimator might be useful for several reasons. Often economic theory does not predict a concrete parametric form, but implies a shape constraint. For example, we may have no reasons that a demand for a good is linear in its price, but it might be plausible to assume that the demand is negatively related to the price, i.e. that it is monotonically decreasing. In that case, imposing a constraint might be necessary to perform policy analysis. Further, imposing constraints on an estimate matters especially in small samples, where deviation from the (assumed) property are likely to occur on random basis. Then, if we believe the constraint must be fulfilled in the population, constrained nonparametric estimators might be the better choice compared to their unconstrained counterparts. Next, opposite to parametric estimates, shape estimates might reveal properties we have not imposed on them. A monotone estimate of a demand function might reveal that the demand is concave. To summarize, on the one hand, constrained nonparametric estimation is more agnostic and flexible than parametric estimation and on the other hand, it might be more useful in small samples than the unconstrained nonparametric estimation.

This section has three main parts. First, I analyze the asymptotic relation between unconstrained and monotonically constrained Tikhonov estimators (section 2.4.3) when the model solution m_0 is an inner point of the constrained set. Next, I derive a projection property of a broad class constrained estimates for a broad class of constraints, see sections 2.4.4 and 2.4.5. And third, this projection property is applied to derive sufficient conditions for the consistency of constrained estimators, see sections 2.4.6 and 2.4.7.

2.4.2 Notation

Unless otherwise stated, G and H represent Hilbert spaces, and $\|.\|_G, \|.\|_H$ the norms on these spaces. T denotes the conditional expectation operator. X, Y, W are one dimensional real-valued random variables defined on some common set. If $\mathscr{A}$ is a sigma field and μ a measure over a set Ω, then $L^p(\Omega, \mathscr{A}, \mu)$ represents the set of μ-measurable p-integrable (with respect to μ functions over Ω. In general, if X is a random variable, then Ω_X is its domain and L^p_X denotes the set of functions $\{\phi : \Omega_X \to \mathbb{R}, \mathbb{E}[\mid \phi(X) \mid^p < \infty]\}$. Denote with $H^k_X := \{m \in L^2(X) : m^{(j)}$ exists and is in $L^2(X)$ for $j = 1 \dots k\}$, where $m^{(j)}$ denotes the jth weak derivative of m (and accordingly with $H^k := H^k_\lambda(\Omega)$ the set $\{m \in L^2(\Omega, \lambda) : m^{(j)}$ exists and is in $L^2(\Omega, \lambda)$ for $j = 1 \dots k\}$). $\|.\|_{L^p}$ denotes the standard norm on L^p, $\|m\| = (\int_\Omega \mid m \mid^p d\mu)^{1/p}$ where the set Ω and the measure μ are for simplicity omitted and taken from the context. Next, if X, Y are random variables (vectors), then their conditional probability functions (c.d.f) are denoted with f_X, f_Y (F_X, F_Y), and $f_{X|Y}(x \mid y)$ ($F_{X|Y}(x \mid y)$) denotes the conditional probability density (c.d.f) of X given $Y = y$ evaluated at $X = x$. If $\{a_n\}$, $\{b_n\}$ are two sequences, then $a_n \asymp b_n$ means that both $\{a_n/b_n\}$ and $\{b_n/a_n\}$ are bounded. I use the big-O little-o notation and the order of probability notation in a standard way. Further notation is introduced where necessary.

2.4.3 Asymptotic equivalence in the case of constraints on the derivatives for a smooth subclass of regression functions

An important question in the context of constraint estimation is whether imposing a constraint has an impact on the convergence rates of the estimator. This question is technically very demanding and is not in the scope of this paper. However, In this subsection, I show that under an additional smoothness condition, imposing constraints on the derivatives in the case of m being an inner point of the constraint monotone set asymptotically leads to the same estimate as in the unconstrained estimation. Thus, the constrained estimator inherits all properties from the unconstrained one. Let in the definition of 2.4.2 $G = H^2_X[0,1]$ and $H = L^2_W[0,1]$, both endowed with their standard norms. Suppose further that the density f_X is bounded and bounded away from zero. Let $C := \{m \in G : m' \leq 0\}$. Further, define the unconstrained and constrained estimators as in section 2.4.1, where $\widehat{T}_n$ and $\widehat{h}_n$ are consistent estimators of T and h and α_n is chosen such that $\|\widehat{m}_n - m_0\|_{H^2_X} = o_p(1)$. Example for such an estimator based on kernels can be found in Grasmair, Scherzer, and Vanhems (2013) and on sieves in Chen and Pouzo (2012). It holds the following result.

Proposition 2.4.1. *Suppose that the model solution m_0 fulfils $m_0 \in G$ and $m_0' \geq c > 0$. Then, as n goes to infinity, it holds*

$$\widehat{m}_n = \widehat{m}_{n,c} \tag{2.4.4}$$

almost everywhere with probability tending to one.

This result can be extended analogically for constraints on higher derivatives, such as convexity. One implication for applied research is that the importance of imposing a constraint on the nonparametric estimator decreases with increasing sample size. One of the purposes of the simulation study in section 2.5 is to provide evidence on how quickly this importance decreases.

Although the smoothness assumption is rather strong, it illustrates well the importance of the properties of the model solution for the relation between constrained and unconstrained estimators.

2.4.4 Constrained Tikhonov estimation as a projection

Simple smooths can be shown to be a projection of the data in some general vector space, Mammen, Marron, Turlach, and Wand (2001). This interpretation has the advantage that constraints are easy to incorporate. In this setting, constrained estimators are revealed to be two-step projection estimators. In addition, constrained smoothing splines are shown to be the projection of the unconstrained counterpart on the set of constrained functions, Mammen and Thomas-Agnan (1999) and Mammen, Marron, Turlach, and Wand (2001). In this and the next subsection, I extend the results of Mammen and Thomas-Agnan (1999) and Mammen, Marron, Turlach, and Wand (2001) to the case of Tikhonov estimators.

Assume $(H, \|\|_H)$, $(G, \|\|_G)$ are two Hilbert spaces, $F : H \to \mathrm{G}$ is a bounded linear operator, h is an element of G representing the (true, perturbated or estimated) data and C is a subset of H. Define $\widehat{m}$ and $\widehat{m}_C$ as

$$\widehat{m} = \underset{m \in H}{\operatorname{argmin}} \|Fm - h\|_H^2 + \alpha_n \|m\|_G^2 \tag{2.4.5}$$

$$\widehat{m}_C = \underset{m \in C}{\operatorname{argmin}} \|Fm - h\|_H^2 + \alpha_n \|m\|_G^2. \tag{2.4.6}$$

As in the previous section, I refer to 2.4.5 and 2.4.6 as to unconstrained and constrained (Tikhonov) estimators, respectively. Existence of $\widehat{m}$ and $\widehat{m}_C$ is assumed throughout this section. Uniqueness is not necessary for my arguments, it only makes notation and exposure easier. Existence and uniqueness are typically closely related to closed and convex constraints.

The following proposition is the central result of this subsection. It gives a useful characterization of the constrained solutions.

Proposition 2.4.4.1. *It holds* $\widehat{m}_C = \operatorname{argmin}_{m \in C} \|F(\widehat{m} - m)\|_H^2 + \alpha_n \|(\widehat{m} - m)\|_G^2$.

This result can be interpreted as follows: the constrained solution $\widehat{m}_C$ is a projection of the unconstrained solution $\widehat{m}$ on the constrained set C. The following corollary makes this interpretation legal.

Corollary 2.4.4.1. *The map* $<,>_\mathscr{V}: H \times H \longrightarrow \mathbb{R}, < m, g >_\mathscr{V} :=< Fm, Fg >_G + \alpha < m, g >_H$ *is for each fixed* α *a positive definite bilinear form. In the induced norm* $\|.\|_\mathscr{V}$ *it holds*

$$\widehat{m}_C = \underset{m \in C}{\operatorname{argmin}} \|\widehat{m} - m\|_\mathscr{V}. \tag{2.4.7}$$

An immediate consequence is the following

Corollary 2.4.4.2. *If the true regression function* m_0 *is in* C*, then for* $0 < K \leq 2$

$$\|\widehat{m}_C - m_0\|_\mathscr{V} \leq K \|\widehat{m} - m_0\|_\mathscr{V}. \tag{2.4.8}$$

If C *is closed and convex, then* $K = 1$*.*

Thus, in the norm $\|\|_\mathscr{V}$, the constrained estimator inherits convergence and convergence rates from the unconstrained estimator. Under the additional assumptions of closedness and convexity of C, $\widehat{m}_C$ converges at least as fast as $\widehat{m}$ towards m_0. This result gives the intuition for the findings in the simulation study in section 2.5. There, in the case of monotonicity constraint, the constrained estimator is shown to outperform its unconstrained counterpart in terms of convergence rates.

Remark: It is desirable to translate these results to the standard norm $\|.\|_H$ on H as typically asymptotic properties are stated with respect to it. In general though the norm $\|.\|_\mathscr{V}$ is weaker than the norm on G. On the one hand , due to $\|m\|_\mathscr{V} \leq (\|F\|(1 + \alpha_n)\|m\|_G)$, so that convergence w.r.t $\|m\|_G$ implies convergence w.r.t $\|m\|_\mathscr{V}$ (the sequence $\{\alpha_n\}$ is a null sequence and is hence bounded). On the other hand, equivalence of the two norms would imply that T is bounded from below and hence its inverse is continuous, which is a contradiction to the chosen setting. Nevertheless, I demonstrate in sections 2.4.6 and 2.4.7 how result 2.4.4.2 can be used to obtain consistency under general constraints in sieves and kernel estimation procedures, respectively.

Remark: Sufficient for proposition 2.4.4.1 is that $\|.\|_G$ and $\|.\|_H$ are induced by symmetric positive semidefinite bilinear forms. In this case, corollary 2.4.4.2 holds with $K = 2$.

2.4.5 Two-Step projection framework

The previous section revealed $\widehat{m}_C$ as the projection of $\widehat{m}$ on the set C with respect to the norm $\|.\|_\mathscr{V}$. The fact that $\widehat{m}$ itself is obtained via minimization leads naturally to the question whether there exists a vector space $(V, \|.\|_V)$, which contains the data vector $Y = (Y_1, \ldots, Y_n)$, the estimators Tikhonov constrained and unconstrained estimators $\widehat{m}$ and $\widehat{m}_C$ respectively in such a way that $\widehat{m}$ coincides with the projection of Y on some subspace U, $\widehat{m} = \operatorname{argmin}_{m \in U} \|Y - m\|_V$ and $\widehat{m}_C$ coincides with the projection of Y on a subset $C \subset U$, $\widehat{m}_C = \operatorname{argmin}_{m \in C} \|Y - m\|_V$, both with respect to the norm $\|.\|_V$. In this setting, the Pythagoras' rule yields

$$\|Y - m\|_V^2 = \|Y - \widehat{m}\|_V^2 + \|\widehat{m} - m\|_V^2, \tag{2.4.9}$$

and thus

$$\widehat{m}_C = \underset{m \in C}{\operatorname{argmin}} \|\widehat{m} - m\|_V, \tag{2.4.10}$$

Thus, the constrained estimator $\widehat{m}_C$ is a two-step projection (TSP) estimator in the following sense: first project the data vector on the space U (unconstrained estimator) and then project this projection on the set of all functions which fulfill the constraint. A two-step projection (TSP) framework was developed in Mammen, Marron, Turlach, and Wand (2001) in the context of simple smooths. In this subsection, I build on their idea to construct a TSP framework in the context of constrained Tikhonov IV estimators.

Sufficient conditions Suppose that $(V, \|.\|)$ is a Hilbert space, that U is a closed subspace of V and that $\mathscr{C}$ is a closed and convex subset of U. Then, using well known relationships from functional analysis, there exists a unique projection $\widehat{m}$ of Y on U and a unique projection $\widehat{m}_c$ of $\widehat{m}$ on $\mathscr{C}$. In what follows, I give a concrete example for $(V, \|.\|_V), U$ and C which fulfill the sufficient conditions.

A concrete TSP framework: Define $H^1 := H^1_\lambda([0,1])$ Consider the vector space V

$$V := \{\vec{m}\,(x) = \begin{pmatrix} m_1(x) \\ m_2(x) \\ \dots \\ m_n(x) \end{pmatrix} : m \in H^1\}.$$

I borrow it from Mammen, Marron, Turlach, and Wand (2001). It contains the subspaces $V_m := \{\vec{m}\,(x) \in V : m_1 = \dots m_n\}$ and $V_{m,c} := \{\vec{m}\,(x) \in V : m_1 = \dots m_n = const\}$. The realization of the data vector $\vec{Y} := (Y_1, \dots, Y_n)^T$ is treated as a vector of constant functions of x. All observations Y_i, X_i, W_i are treated as given (deterministic) realizations. Consider the following bilinear form on V:

$$\begin{aligned} &< ., . >_V : V \times V \longrightarrow \mathbb{R} \\ &(\vec{m}, \vec{g}) \longrightarrow \frac{1}{n} \int (\sum_{i=1}^{n} A_i(w) m_i(X_i))(\sum_{i=1}^{n} A_i(w) g_i(X_i)) dw + \\ &\alpha_n \sum_{i=1}^{n} \frac{1}{n} \int m_i'(x) g_i'(x) dx, \end{aligned}$$

where A_i are nonnegative weights functions. Define $\|\vec{m}\|_V := \sqrt{<\vec{m}, \vec{m}>_V}$. Below I show for the space $(V, \|.\|_V$ that the sufficient condition TSP 1 is fulfilled.

Example With A_i a kernel function, for elements of V_m the expression $\|\vec{Y} - \vec{m}\|^2_V$ corresponds to the functional $\|\widehat{T}m - \widehat{h}\|^2_{L^2} + \alpha_n \|m\|^2_{\mu,H^1}$ in Grasmair, Scherzer, and Vanhems (2013), where $\widehat{T}$ and $\widehat{h}$ are kernel estimators of T and h, respectively and $\|.\|^2_{\mu,H^1}$ is the weighted Sobolev norm $\|m\|^2_{\mu,H^1} = \|m\|^2_{L^2} + \mu\|m'\|^2_{L^2}$ (here with the choice $\mu = 0$). See subsection 2.4.7 for a detailed study of this example.

One technical hurdle in this framework is to show completeness of V with respect to $\|.\|_V$. The crucial aspect is to bound the norm of the function with the norm of the derivative. I achieve this through a modified version of the Poincaré's inequality.

Proposition 2.4.5.1. *The space V is a Hilbert space with respect to the norm $\|.\|_V$. The subspace V_m is closed in V with respect to the topology induced by $\|.\|_V$.*

A direct consequence of proposition 2.4.5.1 is the following

Corollary 2.4.5.1. *(Pythagoras relationship) In $(V, \|\|_V)$ there exists a unique projection $\vec{\widehat{m}}$ of $\vec{Y}$ on V_m. For each $\vec{m} \in V_m$ it holds*

$$\|\vec{Y} - \vec{m}\|_V^2 = \|\vec{Y} - \vec{\widehat{m}}\|_V^2 + \|\vec{m} - \vec{\widehat{m}}\|_V^2. \tag{2.4.11}$$

Note that the projection $\vec{\widehat{m}}$ is a vector-valued function with i-th component equal to $\text{argmin}_{m \in H^1} \|\widehat{T}m - \widehat{h}\|_{L^2}^2 + \alpha_n \|m\|_{\mu,H^1}^2$, $i = 1, \ldots, n$. As the next corollary shows, this representation has the important advantage of a straightforward incorporation of constraints (here with the example of monotonicity constraint). Define $\widehat{m}_C := \text{argmin}_{m \in H^1 : m' \geq 0} \|\widehat{T}m - \widehat{h}\|_{L^2}^2 + \alpha_n \|m\|_{\mu,H^1}^2$

Corollary 2.4.5.2. *(Monotonicity constraint) Let $C \subset V_m$, $C := \{\vec{m} \in V_m : m' \geq 0\}$. Then C is convex and closed in V w. r. t. $\|.\|_V$. Therefore there exists a unique element $\vec{\widehat{m}}_C \in C$ with*

$$\vec{\widehat{m}}_C = \underset{\vec{m} \in C}{\text{argmin}} \|\vec{Y} - \vec{m}\|_V. \tag{2.4.12}$$

It holds the relationship

$$\vec{\widehat{m}}_C = \underset{\vec{m} \in C}{\text{argmin}} \|\vec{\widehat{m}} - \vec{m}\|_V, \tag{2.4.13}$$

and the i-th element of $\vec{\widehat{m}}_C$ is equal to $\widehat{m}_C$, $i = 1, \ldots, n$.

2.4.6 Application of the projection property: an example with a sieves estimator

In this subsection, I give a first illustration of how the projection property can be utilized to show strong convergence of an IV estimator under shape constraints. Convergence is inherited from the unconstrained estimation. The analysis is set in the Penalized Sieve Minimum Distance framework of Chen and Pouzo (2012). In particular, the estimation procedure corresponds to their infinite-dimensional sieve case with a least squares series estimator. In the following, G is a reflexive Hilbert space - either $L^2(X)$ or the Sobolev space of functions $H_X^k := \{m \in L^2(X) : m^{(j)} \quad \text{exists and is in } L^2(X) \text{ for } j = 1 \ldots k\}$, where $m^{(j)}$ denotes the jth weak derivative of m. Let $\{b_j(.)\}$ be a sequence of basis functions in L_W^2. For $J = J(n)$ denote with $\bar{b}^J$ the vector $(b_1(.), ..., b_J(.))'$. Further define $B_J := (\bar{b}^J(W_1), ..., \bar{b}^J(W_n))'$, and with $(B_J'B_J)^-$ the Moore-Penrose pseudoinverse of $B_J'B_J$. Following Chen and Pouzo (2012),

equation 11)) define the Least Squares Estimator of $(Tm - h)(w) = \mathbb{E}[m(X) \mid W = w] - \mathbb{E}[Y \mid W = w]$ as

$$(\widehat{T}m - \widehat{h})(w) := \bar{b}^J(w)(B_J'B_J)^- \sum_{i=1}^{n} \bar{b}^J(W_i)(m(X_i) - Y_i). \tag{2.4.14}$$

Let C some subset of G. Define the unconstrained and the constrained estimators as

$$\widehat{m} \in \underset{m \in G}{\operatorname{argmin}} \{n^{-1} \sum_{i=1}^{n} (\widehat{T}m(W_i) - \widehat{h}(W_i))'(\widehat{T}m(W_i) - \widehat{h}(W_i)) + \alpha_n P(m)\} \tag{2.4.15}$$

$$\widehat{m}_C \in \underset{m \in C}{\operatorname{argmin}} \{n^{-1} \sum_{i=1}^{n} (\widehat{T}m(W_i) - \widehat{h}(W_i))'(\widehat{T}m(W_i) - \widehat{h}(W_i)) + \alpha_n P(m)\}, \tag{2.4.16}$$

where $P(m) = \int m^2(x)dx$ when $G = L^2(X)$ and $P(m) = \sum_{j=0}^{k} \int \mid m^{(j)(x)} \mid^2 dx$ with $m^{(0)} := m$ if $G = H_X^k$ (the domain of X is specified below).

The existence of $\widehat{m}_C$ depends on the properties of the set C. The functional $\mathscr{F}m := n^{-1} \sum_{i=1}^{n} (\widehat{T}m(W_i) - \widehat{h}(W_i))'(\widehat{T}m(W_i) - \widehat{h}(W_i)) + \alpha_n P(m)$ is convex and sequentially lower semi-continuous with respect to $\|.\|$ and hence it is also sequentially lower semicontinuous with respect to the weak topology on G. It is also coercive and bounded from below, and because G is a reflexive space, a sufficient condition for the existence of $\widehat{m}_C$ is that the set C is weakly closed in G, see for example Engl, Hanke, and Neubauer (1996). Also alternative conditions can be found there.

Examples: i) Let $G = L^2(X)$ and let C be the subset of all positive functions, $C = \{m \in H : m \geq 0 \quad a.e.\}$. C is closed and convex and hence weakly closed in G. ii) Let $G = H_X^1$ and let C be the subset of all non-decreasing functions, $C = \{m \in H : m' \geq 0 \quad a.e.\}$. This set is also closed[4] and convex and hence weakly closed. iii)The weak closedness of a constrained set of C is typically obtained through convexity. Grasmair, Scherzer, and Vanhems (2013) give prove weak closedness in the case of the nonconvex integrability constraints of a demand function. The integrability constraints of a demand function $g = g(x)$ for a vector of goods x under prices p and budget z follow from deriving the demand function through a maximization of a utility function u under the budget constraint z. The integrability constraints are defined in the following way: a) g is homogeneous of degree 0: $g(tx) = g(x)$ for every $t > 0$, b) the demand satisfies the equality $p'g(x) = z$, that is, the solution of the utility maximization problem is at the boundary and c) the Slutsky matrix $S_g(x) := \nabla_p g(x) + \partial_z g(x)g(x)'$ is symmetric and negative semidefinite. Grasmair, Scherzer, and Vanhems (2013) show that due to condition c) the set of integrable demand funcitons is nonconvex, see Remark 2.1 in Grasmair, Scherzer, and Vanhems (2013), but is nevertheless weakly closed in the sobolev space $H^1(\mathbb{R}^k)$ with k the dimension of x, see Lemma 2.2 in Grasmair, Scherzer, and Vanhems (2013).

Consider the following collection of assumptions from Chen and Pouzo (2012):

Assumption 1 For m_0 it holds $Tm_0 = h$.

[4]Let the sequence $\{m_n\}$ of functions in C converge to $m \in G$. If m' is negative on a λ-nonnull set M, then the set $\bigcup_{n=}^{\infty} A_n$ with $A_n := \{m' \leq -1/n\}$ has a positive measure and therefore at least one of the sets A_n must have a positive measure, which is a contradiction to the convergence.

Assumption 2 T is bounded and if for some m $Tm = h$ then $\|m - m_0\|_G = 0$.

Assumption 3 i) X and W attain values on $[0,1]$. ii) The densities f_X and f_W of X and W are bounded and bounded away from zero. iii) The joint distribution of (X,W) is dominated by the products of its margins and $\int \mid f_{X,W} \mid^2 df_x \otimes f_y < \infty$. iv) Y is square integrable. v) $(Y_i, X_i, W_i), i = 1,2,\ldots,n$ is an i.i.d. random sample from the distribution of (Y,X,W).

Assumption 4 i) $sup_{j\in\{1,\ldots,J_n\}}\mathbb{E}[\mid b_j(X) \mid^2] \leq C_J$, where $\{C_J\}_{J\in\mathbb{N}}$ is a sequence of positive constants. ii) The smallest eigenvalue of $\mathbb{E}[\bar{b}^{J_n}(W)\bar{b}^{J_n}(W)']$ is bounded away from zero for all $J_n \in \mathbb{N}$. iii) $\sup_{w\in[0,1]}\{b_1^2(w) + \cdots + b_{J_n}^2(w)\} = o(n/J_n)$.

Remark: Denote with $\widehat{m}_n = \widehat{m}(n)$ the unconstrained estimator 2.4.15 corresponding to a given sample size n. Assume that $n^{-1}\sum_{i=1}(\widehat{T}m_0 - \widehat{h})^2 = O_p(\alpha_n)$. By definition of $\widehat{m}_n$

$$n^{-1}\sum_{i=1}(\widehat{T}\widehat{m_n} - \widehat{h})^2 + \alpha_n P(\widehat{m}_n) \leq n^{-1}\sum_{i=1}(\widehat{T}m_0 - \widehat{h})^2 + \alpha_n P(m_0),$$

and therefore exists a constant C_0 such that $\alpha_n P(\widehat{m}_n) \leq \alpha_n C_0$. Note that the assumptions of Lemma A.3 ii) in Chen and Pouzo (2012) imply even $P(\widehat{m}_n) = P(m_0) + o_p(1)$. This justifies the definition of the following set. Define for some arbitrary small $\epsilon > 0$ the set $H^{M_0} := \{m \in G : \alpha_n P(m) \leq \alpha_n M_0\}$ where $0 < M_0 < \infty$ is chosen so that $m_0 \in H^{M_0}$ and $P\{\widehat{m}_n \in H^{M_0}\} \geq 1 - \epsilon$ for all $n \geq n_0$ (n_0 depends on ϵ).

Assumption 5 There exist constants $k \in (0,1]$ and $0 < K < \infty$, so that

$$\max_{1\leq j\leq J_n}\mathbb{E}[(b_j(W))^2 sup_{m'\in H^{M_0}:\|m'-m\|_G\leq\delta}\{(m'(X) - m(X))^2\}] \leq K^2\delta^{2k} \tag{2.4.17}$$

for all $m \in H^{M_0}$ and all $\delta > 0, \delta = o(1)$.

Discussion of Assumptions: Assumption 1 states the model equation. Assumption 2 is maintained for (global) identification of m_0. Under Assumption 2, the conditional expectation operator T is injective and m_0 is the unique solution of the model a.e.. The assumption corresponds to the notion of complete statistic, see Darolles, Fan, Florens, and Renault (2011) for a discussion. Assumption 3 iii) is Assumption A.1 from Darolles, Fan, Florens, and Renault (2011). It ensures that T, its adjoint operator T*, as well as TT* and T*T are compact. Necessary condition for it to hold is that W and X have no elements in common (or in the one dimensional case, that $X \neq W$). Assumption 3 iv) is a sufficient condition that $h = \mathbb{E}[Y \mid W]$ is in $L^2(W)$. Assumption 3 v) can be relaxed to (Y_i, X_i, W_i) strictly stationary ergodic. Assumption 5) is first used in Chen, Linton, and Keilegom (2003). It is an extension of the type IV class functions assumption in Andrews (1994). In Chen, Linton, and Keilegom (2003) it is referred to as locally uniformly $L^2(W)$-continuity. This condition has been used also by Breunig (2013). Further, under assumptions 1-5, there are positive finite numbers K, K', such that

$$K\mathbb{E}[|Tm(W) - h(W)|^2] - O_p(\delta_n^2) \leq n^{-1}\sum_{i=1}^{n}(\widehat{T}m(W_i) - \widehat{h}(W_i))^2 \tag{2.4.18}$$
$$\leq K\mathbb{E}[|Tm(W) - h(W)|^2] + O_p(\delta_n^2)$$

uniformly over the set $\mathscr{H}_{os}$ which is defined below, where $\delta_n^2 = \max\{J_n/n, b_{J_n}^2\}$ and b_{J_n} is the bias of the series estimator, see lemma 3.2 Chen and Pouzo (2012).

In the following result, I impose a high-level condition, under which the consistency of the unconstrained estimator implies the consistency of the constrained estimator. This relation is a consequence of the projection property of the constrained estimator. Note that the high-level condition is not a direct condition on the set C.

Proposition 2.4.2 (Consistency)**.** *Let m_C exist and assumptions 1-5 be fulfilled. If*

$$n^{-1}\sum_{i=1}^{n}(\widehat{T}(\widehat{m}-m_0))^2 = o_p(\alpha), \tag{2.4.19}$$

then $\|\widehat{m}_C - m_0\|_G = o_p(1)$

I now show that the high-level assumption 2.4.19 is implied by low-level conditions that are directly related to the smoothness of the model solution m_0 and the rate of ill-posedness of the inverse problem. We first need to introduce the following notation and assumptions. Define $\mathscr{H}_{os}$ as $\mathscr{H}_{os} := \{m \in G : \|m - m_0\|_G \leq \epsilon, \|m\|_G \leq M_1\}$ for $M_1 > 0$ and ϵ such that that $P\{\widehat{m_0} \in \mathscr{H}_{os}\} > 1 - \epsilon$ and $m_0 \in \mathscr{H}_{os}$. Note that due to the definitions of G and $\|.\|_G$, this set is convex and assumption 4.1 of Chen and Pouzo (2012). Let $\{\delta_n^2\}$ be a sequence of positive decreasing numbers such that $\delta_n^2 = \max\{J_n/n, b_{J_n}^2\}$ and b_{J_n} and 2.4.18 holds uniformly over $\mathscr{H}_{os}$(Such sequence exists due to assumptions 1-5, see Lemma C.2 in Chen and Pouzo (2012)). Further, define $\delta_{m,n}^*$ to be the optimal convergence rate of $\mathbb{E}[|\, Tm - h - (\widehat{T}m - \widehat{h})\,|^2]$ uniformly over $\mathscr{H}_{os}$.

Assumption 6 $\max\{\delta_n^2, \alpha_n\|\widehat{m} - m_0\|_G\} = O(\delta_n^2)$.

Under assumptions 1-6, $\|\widehat{m} - m_0\|_G = O_p((\delta_n^2, \mathscr{H}_{os}))$, where

$$(\delta_n^2, \mathscr{H}_{os}) := \sup_{m\in\mathscr{H}_{os}:\mathbb{E}[|\mathbb{E}[m(X)-m_0(X)|W]|^2]} \|m - m_0\|_G \tag{2.4.20}$$

is a measure of ill-posedness of T called the sieve modulus of continuity, see Blundell, Chen, and Kristensen (2007). Let $\{q_j(.)\}j \in \mathbb{N}$ be a Riesz basis functions in G. The following assumption states the rate of the approximation error of the projection on subspace spanned by finite many basis functions (this is assumption 5.3 Chen and Pouzo (2012)).

Assumption 7 There exist finite constants $M > 0$ and $\alpha > 0$, and a strictly increasing positive sequence $\{v_j\}_{j\in\mathbb{N}}$, such that $\|m - \sum_{j=1}^{k}\langle m, q_j\rangle_G q_j\|_G \leq M(v_{k+1})^{-\alpha}$.

Assumption 8 There exist a finite constants $D > 0$ and a continuous strictly increasing function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$, such that for all m in $\mathscr{H}_{os}$, it holds $\sqrt{\mathbb{E}[|\, \mathbb{E}[m(X) - m_0(X) \mid W]\,|^2]} \leq D\sum_{j=1}^{\infty}\phi(v_j^{-2}) \mid \langle m - m_0, q_j\rangle_G \mid^2$.

Assumptions 7 and 8 are related to the smoothness of the model solution m_0 and to the rate of ill-posedness of the inverse problem. Combined, they provide an upper bound for the sieve modulus of continuity, see Chen and Pouzo (2012) corollary.

Note that assumptions 1-8 are used to derive the convergence rates of the unconstrained estimator in a Hilbert space. Provided that m_0 is in C, they are only indirectly related to the constrained set C. The following result states, that provided that the estimation problem is enough well-behaved, no further assumptions on C are necessary to ensure consistency of the constrained estimator (apart from its existence).

Proposition 2.4.3. *Suppose that assumptions 1-8 hold and that the inverse problem is mildly ill-posed in the sense, that $\phi(x) = x^s$ for some $s > 0$. In addition, assume that*

(i) $\alpha - 2s > 0$

(ii) $T\widehat{m} - Tm_0$, $\widehat{T}m_0 - Tm_0$ *and* $\widehat{h} - h$ *suffice the uniform law of large numbers.*

Then 2.4.19 holds.

Ceteris paribus, α controls the rate of the approximation error on the shrinking set. The function $\phi(x)$ links the behaviour of the conditional expectation operator to the smoothness of the model solution. Assumption 8 i) can be therefore interpreted as a local ill-posedness condition. Condition ii) can be ensured by assuming that there exist numbers r, q and measurable square integrable such that $Tm_0 \in H_W^r$ and $\widehat{h} \in H_W^q$ and that the consistent estimators $\widehat{T}m_0$, $T\widehat{m}$ and $\widehat{h}$ are bounded there for sufficiently large n by an envelope function, see e.g. Van de Geer (2000).

2.4.7 Application of the projection property: an example with a kernel estimator

In this subsection, I utilize the projection property in a kernel based estimation approach. As in the last example, I state an assumption, under which convergence of the unconstrained estimator implies the convergence of the constrained estimator, provided it exists, regardless of the type of constraint. As before, I state sufficient conditions for this assumption to hold. These conditions are related to the smoothness of the true function and the ill-posedness of the problem.

Using the notation of the projection section 2.4.4, let $G = L_X^2([0,1])$, $H = L_W^2([0,1])$, where both spaces are endowed with their standard norms. To define estimators, first consider the following definition of a generalized kernel function (compare Darolles, Fan, Florens, and Renault (2011)).

Let $\sigma = \sigma(n) \to 0$ and define the function $K_\sigma(.,.)$ with the following properties:

i) $K_\sigma(u,t) = 0$ for $u > t$ and $u < t - 1$,

ii) $\sigma^{-1} \int_{t-1}^{t} K_\sigma(u,t)du = 1$,

iii) $\sigma^{-1+j} \int_{t-1}^{t} u^j K_\sigma(u,t)du = 0$ for $j = 1, 2 \ldots, l-1$

K_σ is called a univariate generalized kernel of order l and σ is the bandwidth. The multivariate generalized kernel of order l is defined as a product of univariate generalized kernels of order

l. The kernel estimators of the densities are defined as

$$\widehat{f_{YW}}(y,w) = \frac{1}{n\sigma^2}\sum_{i=1}^{n} K_{Y,\sigma}(y-Y_i,y)K_{W,\sigma}(w-W_i,w),$$
$$\widehat{f_{XW}}(x,w) = \frac{1}{n\sigma^2}\sum_{i=1}^{n} K_{X,\sigma}(x-X_i,x)K_{W,\sigma}(w-W_i,w),$$
$$\widehat{f_{W}}(w) = \frac{1}{n\sigma}\sum_{i=1}^{n} K_{W,\sigma}(w-W_i,w),$$

while the estimators of T and r are

$$\widehat{T}m(w) = \int m(x)\frac{\widehat{f_{XW}}(x,w)}{\widehat{f_{W}}(w)}dx,$$
$$\widehat{r}(w) = \int y\frac{\widehat{f_{YW}}(y,w)}{\widehat{f_{W}}(w)}dx,.$$

The functional to be minimized over C is now defined as

$$\widehat{\mathscr{F}_n}(m) := \int \mid \widehat{T}m(w) - \widehat{r}(w) \mid^2 \lambda(dw) + \alpha_n \int \mid m(x) \mid^2 \lambda(dx).$$

The unconstrained and constrained estimators are defined as $\widehat{m} = \text{argmin}_{m\in L_X^2}\,\widehat{\mathscr{F}_n}(m)$ and $\widehat{m}_C = \text{argmin}_{m\in C}\,\widehat{\mathscr{F}_n}(m)$, where C denotes some subset of L_X^2. The following collection of assumptions is borrowed from Darolles, Fan, Florens, and Renault (2011).

Assumption K_1
1) $(Y_i, X_i, W_i), i1 = 1, \ldots, n$ is an i.i.d. sample of (Y, X, W).
2) The density $f_{Y,X,W}$ is l times continuously differentiable in the interior of its domain and $f_{X,W}$ is bounded and bounded away from zero on $[0,1]\times[0,1]$.
3) $\mathbb{E}[\epsilon \mid W = w]$ is uniformly bounded on [0,1]

Assumption K_2
The operator T is an injective bounded linear operator.

The i.i.d. assumption can be relaxed to weak stationarity. Assumptions K_1 and K_2 ensure that the operator T is compact with a single value decomposition. Denote it with $\{\mu_k, e_k, f_k\}_{k\in\mathbb{N}}$, where (μ_k) is a decreasing positive sequence with $\mu_k \to 0$ and (e_k) and (f_k) are sequences of the corresponding eigenfunctions in $L^2(X)$ and $L^2(W)$, respectively. The injectivity of T ensures global identification of m_0.

Assumption K_3
The true regression function m_0 is in C.

Define the following class of functions. For $\beta > 0$, denote with Φ_β the set of functions m in G such that $\sum_{i=1}^{\infty}\langle m, e_i\rangle_G^2/\mu_i^{2\beta} < \infty$

Assumption K_4
For some $\beta > 0$ it holds $m_0 \in \Phi_\beta$.

Assumption K_4 is called a source condition and is related to both the (decay of) singular values of T and the (decay of) the Fourier coefficients of the function m_0, see Darolles, Fan, Florens, and Renault (2011) and Engl, Hanke, and Neubauer (1996) for a discussion. The parameter β regulates this relation, ceteris paribus.

Assumption K_5
The kernel function satisfies the following properties:
1) K_σ is a univariate generalized kernel of order l.
2) For $t \in [0,1]$ $\text{supp}(K_\sigma(.,t)) = [(t-1)/\sigma, t/\sigma] \cap \mathscr{K}$, with $\mathscr{K}$ not depending on t. Moreover, $\sup_{\sigma>0, t\in[0,1], u\in\mathscr{K}} \mid K_\sigma(\sigma u, t) \mid < \infty$.
3) It holds $\log(n)/(n\sigma^4) \to \infty$ and $\sigma \to \infty$.

Remark Assumptions $K_1 - K_5$ are sufficient to ensure the consistency of the unconstrained estimator under suitable choices of the bandwidth and the penalty term, see Darolles, Fan, Florens, and Renault (2011). Moreover, under such choices of the bandwidth and the penalty term, there exists $\rho \leq 2$ such that $\|\widehat{T} - T\|^2 = O_p(1/n\sigma^2 + \sigma^{2\rho})$, $\|\widehat{T}^* - T^*\|^2 = O_p(1/n\sigma^2 + \sigma^{2\rho})$, and $\|\widehat{T}^*\widehat{r} - \widehat{T}^*\widehat{T}m_0\| = O_p(1/n + \sigma^{2\rho})$, where T^* denotes the adjoint of T and $\|\|$ is the operator norm, see the supplement materials to Darolles, Fan, Florens, and Renault (2011). Note that, apart from condition K_3, none of the conditions $K_1 - K_5$ is directly related to the constrained set. Assumption K_4 is implicitely related to C, as it states that $m_0 \in \Phi_\beta$ is also in C.

The following proposition states a sufficient condition that the constrained estimator, provided it exists, is a consistent estimator for m_0 with respect to the norm on G.

Proposition 2.4.4. *Let assumptions K1-K6 be fulfilled and let*

$$\int \mid \widehat{T}(\widehat{m}(w) - m_0)(w) \mid^2 \lambda(dw) = o_p(\alpha_n) \tag{2.4.21}$$

hold. Furthermore, let i) $\alpha_n = o(1)$ *with* $n\alpha^2 \to \infty$, *ii)* $\sigma = o(1)$ *with* $n\sigma^2 \to \infty$ *and iii)* $\beta \geq 1$ *or* $n\sigma^2\alpha_n^{1-\beta} \to \infty$. *Then, if* $\widehat{m}_C$ *exists, it holds*

$$\|\widehat{m}_C - m_0\|_G = o_p(1).$$

The intuition for the result is similar as the consistency result for constrained sieves estimators from section 2.4.6. Under the assumptions of proposition 2.4.4, the unconstrained estimator is consistent and hence $\alpha_n\|\widehat{m}_C - m_0\|_G^2 = o_p(\alpha_n)$. Together with the condition 2.4.21, the projection property 2.4.4.2 ensures the consistency of the constrained estimator.

In the following I discuss sufficient conditions for the assumption 2.4.21 to hold. As in the sieves application, a condition on the smoothness of m_0 and on the rate of the ill-posedness of the inverse problem are revealed to imply it.

Proposition 2.4.5. *Let assumptions $K_1 - K_5$ hold and let $\widehat{m}_C$ exist. If i) $\alpha_n = o(1)$ with $n\alpha^2 \to \infty$, ii) $\sigma = o(1)$ with $n\sigma^2 \to \infty$ and iii) $\beta > 1$, then under optimal choices of the penalty and the bandwidth it holds*

$$\int | \widehat{T}(\widehat{m}(w) - m_0)(w) |^2 \lambda(dw) = o_p(\alpha_n)$$

2.4.8 Discussion

The results of sections 2.4.6 and 2.4.7 imply, that convergence of the constrained estimator is implied by conditions on the smoothness of the model solution m_0 and the rate of ill-posedness of the inverse problem. For sieves, this condition was $\alpha - 2s > 0$ and for kernels $\beta > 1$. The parameter s controls the degree of ill-posedness in the sieves case. The higher the s, the higher the rate. α is related to the sieve approximation error. The higher α, the higher the rate at which the approximation error goes to zero with increasing dimension of the sieve space. It can be interpreted as a smoothness requirement. A similar relation holds for the parameter β in the kernel case. The case $\beta > 1$ is interpreted in Darolles, Fan, Florens, and Renault (2011) as a strong instrument case.

One possibility to use further the projection property would be to bound the convergence rates of the constrained estimators. Suppose that $\{R_n\}$ is some positive decreasing null sequence, and that $\|\widehat{m} - m_0\|_G = o_p(R_n)$. In the sieves case, assume that

$$n^{-1} \sum_{i=1}^{n} (\widehat{T}(\widehat{m} - m_0))^2 = O_p(\alpha_n \|\widehat{m} - m_0\|_G), \tag{2.4.22}$$

and in the kernel case, assume that

$$\int | \widehat{T}(\widehat{m}(w) - m_0)(w) |^2 \lambda(dw) = O_p(\alpha_n \|\widehat{m} - m_0\|_G). \tag{2.4.23}$$

In both cases, using the projection property, it would hold $\|\widehat{m}_C - m_0\|_G = o_p(R_n)$. It is not in the scope of this paper to discuss the plausibility of assumptions 2.4.22 and 2.4.23, as well as sufficient conditions for them to hold.

2.5 Simulation study and a guideline for the applied research

2.5.1 Estimation

The simulation study in this section has several purposes. First, it seeks to determine rules for choosing the estimation parameters in an optimal way. The choice will generally depend on the underlying functional form (smoothness and curvature of the model solution m_0), as well as on the rate of the ill-posedness of the inverse problem. The latter property is related to the strength of the instrument and to the level of endogeneity in the model. Therefore, I simulate (i) different functional forms, (ii) different strengths of the instrument and different level of endogeneity for (iii) different sample sizes. This framework allows me to formulate guidelines for applied research for the choice of the estimation parameters when imposing a constraint is necessary.

A second purpose of this simulation study is to compare the performances of unconstrained and constrained estimators and relate them to the theoretical findings from the previous sections. Of particular interest is whether there are statistical gains from imposing a (correctly specified) constraint. I seek to answer the question how these gains change with increasing n depending on the structure of the model (strength of the instrument, smoothness of the solution).

The preceding sections in this paper delivered some insights and predictions about the behavior of the constrained estimator.

1. Due to the projection property, the constrained estimators should perform at least as good as their unconstrained counterparts.

2. When m_0 is an inner point of the constrained set, then constrained and unconstrained estimators are equal w.p.t.1.

I am using the PSMD sieves estimator discussed in section 2.4.6. Here m_n denotes the number of bias functions used to compute the LS series estimator. The main determinant of the optimal values of the parameters of the estimator is the functional form of the true regression. In order to cover a wide range of functional forms, I simulate three models that have different "degree" of monotonicity: i) a strictly increasing function (model 1) with $m_1(x) = (\exp(x))^2$; ii) a model that has a strictly increasing and a flat part (model2):

$$m_2(x) = \begin{cases} 0.5x - 0.5 & \text{if} \quad x \in [-1, -0.7] \\ -0.625x^2 - 0.375x + 0.19375 & \text{if} \quad x \in [-0.7, -0.3] \\ 0.25 & \text{if} \quad x \in [-0.3, 0.3] \\ 0.625x^2 - 0.375x + 0.30625 & \text{if} \quad x \in [0.3, 0.7] \\ 0.5x & \text{if} \quad x \in [0.7, 1]; \end{cases}$$

and iii) a constant function, $m_3 = 0$. Fugures 2.1a- 2.2 depict the three models.

Which is the most plausible form will depend on the available theory.

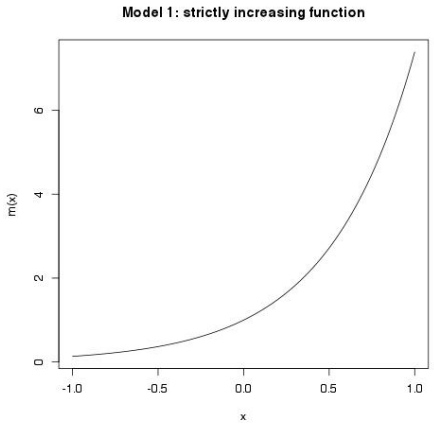
Two further important factors for choosing the parameters of the estimator are the degree of endogeneity and the strength of the instrument. Whereas the strength of the instrument can be directly assessed, the degree of the endogeneity is unobservable. To address this issue, we can rely mainly on economic arguments and indirect evidence. I simulate different combinations of strength of the instrument and degree of endogeneity in the following way. The endogenous regressor X is generated according to the rule

$$X_{k,i} = \beta_1 \epsilon_i + \beta_2 W_{u,i} + \beta_3 U_i,$$

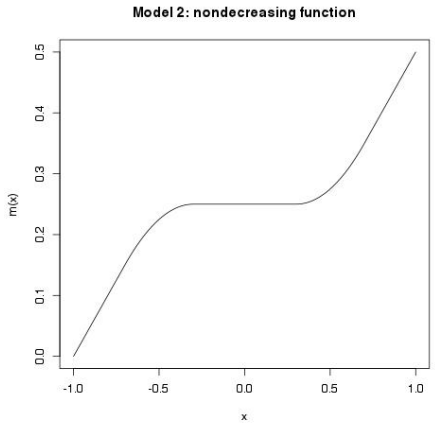
where ϵ_i is the regression error and U_i is exogenous variation of X which is independent of the instrument. W denotes the instrument. The constant β_1 controls the degree of endogeneity, β_2 controls the strength of the instrument, and β_3 controls the exogenous variation of X. β_1, β_2 and β_3 are nonnegative and sum up to one. For details of the data generation process see the appendix. For each functional form I simulate three different degrees of endogeneity: strong endogeneity- weak instrument (SE-WI), weak endogeneity-strong instrument (WE-SI), weak endogeneity-weak instrument (WE-WI), compare with table 2.2.
This data generating process has three main advantages. First, the variety of models allows a

Figure 2.1: Simulated models

(a) Model 1



(b) Model 2



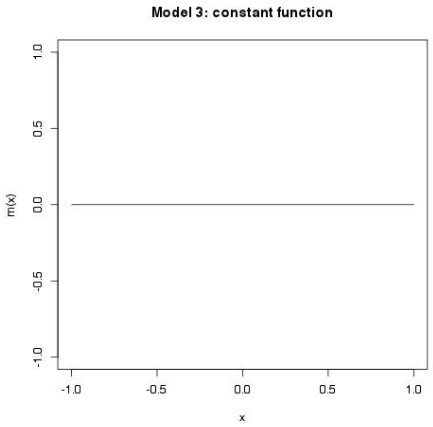


Figure 2.2: Model 3

Table 2.2: Overview case definitions

	β_1	β_2
SE-WI	0.5	0.2
WE-SI	0.2	0.5
WE-WI	0.2	0.2

thorough investigation of the influence of the endogeneity and the strength of the instrument.

Second, the simplicity of the model facilitates the interpretation of the results. And third, the values of these two factors are easy to manipulate and adjust according to our theory. The main questions I will focus on are the following:

- How quickly converges the constrained estimator with increasing sample size compared to the unconstrained estimator (MISE and uniform convergence)?
- What is the influence of endogeneity on this convergence?
- What are the optimal parameters (cutoff parameter J_n and regularization constant α_n for a given functional form of the true function, endogeneity level and sample size? How do they change when these given conditions change?

The three parameters of the Monte Carlo simulation are the sample size n, the cutoff parameter J_n and the regularization constant α_n. They can attain the following values:

$$n \in \{50, 100, 300, 500, 700, 1000\}$$
$$J_n \in \{3, 4, 5, 6, 7, 8, 9, 10\}$$
$$\alpha_n \in \{0.0001, 0.0003, 0.0005, 0.0008, 0.001, 0.003, 0.005, 0.008, 0.01, 0.03, 0.08, 0.1\}.$$

For each model (1,2,3) and for each triple $(\beta_1, \beta_2, \beta_3)$ (in total 9 submodels) I simulate $N = 100$ Monte Carlo samples for each triple (n, J_n, α_n) and calculate the empirical counterpart of MISE.

Comparison of the performance of constrained and unconstrained estimators For each of the 9 submodels and for each sample size, the MISE-minimizing parameters m and α_n for both constrained and unconstrained estimators are determined and the minimal MISEs are compared. Tables 2.3 - 2.5 summarize the results for the MISE. A comparison between the performances of the constrained and unconstrained estimator reveals that the unconstrained estimator outperforms the unconstrained one for each model and sample size, although the differences disappear when n gets bigger. Figures 2.3 - 2.5 illustrate these findings. In the case of model 3, the estimation amounts to simple averaging and the two procedures produce equivalent results. As expected, the performance of the estimators get worse with rising endogeneity and decreasing strength of the instrument.

Optimal parameters and guidelines for the applied research

As noted, the choice of a parameter set in a concrete situation should be made according to the sample size and according to the theory about 1) the functional form of the true regression function and 2) the degree of endogeneity and the strength of the instrument in the data. The following observations can be made about the optimal cutoff parameter J_n (see tables 2.6, 2.7 and 2.8):

- it grows very slowly with the sample size.
- the higher the degree of endogeneity, the higher J_n
- the stronger the instrument, the smaller J_n
- (trivially) the more "complex" the functional form the higher J_n

The following observations can be made about the optimal regularization parameter α_n (see tables 2.9, 2.10 and 2.11):

Table 2.3: Simulated MISE WE-SI

	Model 1		Model 2		Model 3	
n	constr.	unconstr.	constr.	unconstr.	constrained	unconstr.
50	229	354	38	42	22	25
100	134	240	27	29	10	11
300	60	90	14	15	3.6	4.2
500	40	54	11.13	11.73	2.1	2.35
700	34	44	8.95	10.15.	1.32	1.63
1000	25	33	6.46	8.7	1.1	1.2

Table 2.4: Simulated MISE SE-WI

	Model 1		Model 2		Model 3	
n	constr.	unconstr.	constr.	unconstr.	constrained	unconstr.
50	835	1014	37	38	21	21
100	702	805	24	25	11	11
300	474	738	15	21	3	3
500	472	649	13	15	2	2
700	411	591	12	14	1	1
1000	408	533	9	12	0.8	0.84

Table 2.5: Simulated MISE WE-WI

	Model 1		Model 2		Model 3	
n	constr.	unconstr.	constr.	unconstr.	constrained	unconstr.
50	742	925	38	44	25	26
100	551	839	27	35	12	12
300	559	700	16	20	3	4
500	479	659	14	17	2	2
700	474	648	10	14	1	1
1000	395	591	9	11	0.9	0.9

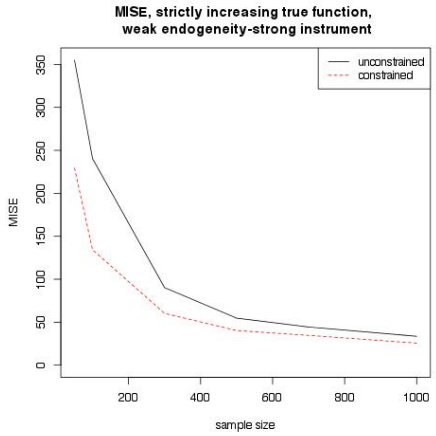


Figure 2.3: MISE Weak endogeneity - strong instrument, Model 1

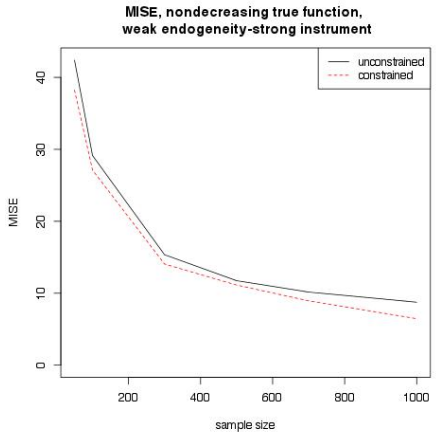


Figure 2.4: MISE Weak endogeneity - strong instrument, Model 2

- it decreases with increasing sample size, as predicted by the theory
- the higher the degree of endogeneity, the higher α_n
- the stronger the instrument, the lower α_n
- the closer the true function to the boundary, the higher α_n.

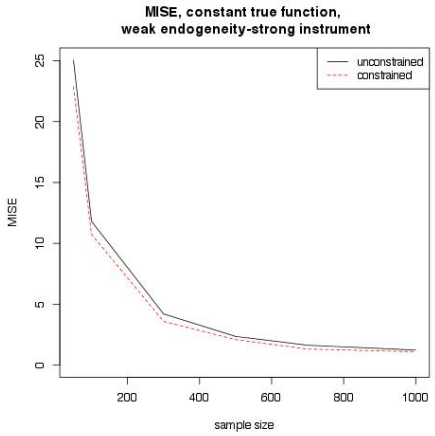


Figure 2.5: MISE Weak endogeneity - strong instrument, Model 3

Table 2.6: Optimal J_n for the constrained estimator, Model 1

n	SE-WI	WE-SI	WE-WI
50	9	4	7
100	8	4	9
300	8	4	5
500	5	4	7
700	7	4	8
1000	7	4	6

Table 2.7: Optimal J_n for the constrained estimator, Model 2

n	SE-WI	WE-SI	WE-WI
50	6	3	8
100	5	3	8
300	10	3	9
500	6	4	3
700	10	4	9
1000	4	4	9

Table 2.8: Optimal J_n for the constrained estimator, Model 3

n	SE-WI	WE-SI	WE-WI
50	3	4	3
100	4	3	3
300	5	5	7
500	4	4	3
700	3	5	3
1000	4	8	4

Table 2.9: Optimal α_n for the constrained estimator, Model 1

n	SE-WI	WE-SI	WE-WI
50	0.003	0.0005	0.0008
100	0.0008	0.0003	0.0003
300	0.0001	0.0003	0.0001
500	0.0001	0.0001	0.0001
700	0.0001	0.0001	0.0001
1000	0.0001	0.0001	0.0001

Table 2.10: Optimal α_n for the constrained estimator, Model 2

n	SE-WI	WE-SI	WE-WI
50	0.01	0.01	0.1
100	0.08	0.01	0.03
300	0.03	0.03	0.008
500	0.01	0.0008	0.008
700	0.01	0.0003	0.008
1000	0.0001	0.0003	0.008

Table 2.11: Optimal α_n for the constrained estimator, Model 3

n	SE-WI	WE-SI	WE-WI
50	0.1	0.1	0.1
100	0.1	0.1	0.1
300	0.1	0.1	0.1
500	0.1	0.1	0.1
700	0.1	0.1	0.1
1000	0.1	0.1	0.1

2.5.2 Testing

Null hypothesis and an empirical testing procedure

In this section, I develop an empirical testing procedure for testing the null hypothesis

$$H_0 : m_0' \geq 0 \quad \text{a. e. on } [0,1]. \tag{2.5.1}$$

The alternative is that $m' < 0$ on a set with positive measure. To construct an empirical test, I use the test developed by Breunig (2012) and Breunig (2013) and adapt it to the case of monotonicity constraint. The intuition of the procedure is as follows. The test is based on an empirical distance between an estimator that is subject to monotonicity constraint and the data. Suppose that $\widehat{m}_C$ is a nonparametric estimator, such that $\widehat{m}_C' \geq 0$ and $\|\widehat{m}_C - m_0\|_S \to 0$ with $n \to \infty$ whenever m_0 is itself monotone ($\|.\|_S$ is some norm on the underlying space). Denote with Y the data and with $d(m, Y)$ some empirical distance between the function m and the data Y. The idea of the test is to use a distance function, such that $d(m, Y)$ is small with a high probability when m_0 is monotone. Such a comparison potentially faces the hard task to derive the distribution of the constrained estimator (or of a function of it). I avoid this problem in the following way. Consider the instrumental equation $\mathbb{E}[Y \mid W] = \mathbb{E}[m_0(X) \mid W]$. If $\widehat{m}_C$ is consistent under the null, then the distance between $\widehat{\mathbb{E}}[Y \mid W]$ and $\widehat{\mathbb{E}}[\widehat{m}_C \mid W]$, $d(Y, \widehat{m}_C) := \|\widehat{\mathbb{E}}[Y \mid W] - \widehat{\mathbb{E}}[\widehat{m}_C \mid W]\|^2$, is likely to be small under the null. Using the binomial formula, the last expression can be split into

$$\begin{aligned} &\|\widehat{\mathbb{E}}[Y \mid W] - \widehat{\mathbb{E}}[m_0 \mid W]\|^2 + 2\langle \widehat{\mathbb{E}}[Y - m_0 \mid W], \widehat{\mathbb{E}}[m_0 - \widehat{m}_C \mid W] \rangle + \\ &\|\widehat{\mathbb{E}}[m_0 \mid W] - \widehat{\mathbb{E}}[\widehat{m}_C \mid W]\|^2. \end{aligned} \tag{2.5.2}$$

If $\widehat{m}_C$ converges quickly enough to m_0, then the distribution of 2.5.2 is determined by its first term. This idea is used in Horowitz (2006) and Breunig (2012) for testing for a parametric specification and in Horowitz (2012), Breunig (2012) and Breunig (2013) for testing of the existence of a smooth regression function in different settings. My contribution is to modify the test procedure in Breunig (2012) to empirically test for monotonicity.

In particular, suppose that (f_j) is a sequence of functions on H such that

$$\int_H f_j(w) f_k(w) \pi(dw) = \begin{cases} 1 & \text{if} \quad j = k \\ 0 & \text{else.} \end{cases}$$

Further, define the infinite dimensional matrix Ω to have an ijth-element the quantity $\mathbb{E}[\epsilon^2 f_i(W) f_j(W)]$, and Ω_{J_n} to be its upper $J_n \times J_n$ block. Define $b_{J_n} = tr(\Omega_{J_n})$ and $\sigma_{J_n} = \sqrt{\Omega_{J_n}^t \Omega_{J_n}}$, with A^t the transposed of a matrix A. In Breunig (2012), the (standardized) test statistic is defined as

$$\mathscr{T}_n = \frac{n S_n - b_{J_n}}{\sqrt{2}\sigma_{J_n}}, \tag{2.5.3}$$

where $S_n = \sum_{j=1}^{J_n} \left[\frac{1}{n} \sum_{i=1}^n (Y_i - \widehat{m}_C(X_i)) f_j(W_i) \right]^2$ and $\widehat{m}_C$ is a nonparametric smooth estimator.

The idea of Breunig (2012) is to split S_n in the following way

$$S_n = \sum_{j=1}^{J_n} \Big(n^{-1} \sum_{i=1}^{n} (Y_i - m_0(X_i)) f_j(W_i) \Big)^2 +$$
$$+2 \sum_{j=1}^{J_n} \Big((n^{-1} \sum_{i=1}^{n} (Y_i - m_0(X_i)) f_j(W_i)) (n^{-1} \sum_{i=1}^{n} (m_0(X_i) - \widehat{m}_C(X_i)) f_j(W_i)) \Big) +$$
$$+ \sum_{j=1}^{J_n} \Big(n^{-1} \sum_{i=1}^{n} (m_0(X_i) - \widehat{m}_C(X_i)) f_j(W_i) \Big)^2 = I_n + II_n + III_n$$

and to show that if the model solution is sufficiently smooth, then all terms containing the nonparametric estimator (II_n and III_n) are asymptotically negligible in the sense that $nII_n = o_p(\sigma_{J_n})$ and $nIII_n = o_p(\sigma_{J_n})$. The distribution of the test statisic is then dominated by nI_n, which behaves like the error term and is normally distributed under the null.

My empirical approach is to take $\widehat{m}_C$ to be a monotonically constrained estimator. A necessary condition is that $\widehat{m}_C$ converges quickly to m_0 under the null. I adapt the penalized sieve minimum distance (PSMD) estimator by Chen and Pouzo (2012) to the case of of monotonicity shape restriction. This estimation procedure is discussed in section 2.4.6. A crucial observation is that because the constrained set is closed and convex in the original space, the constrained estimator achieves the convergence rate bound derived for the unconstrained counterpart. See corollary 5.2 in connection with theorem 4.1 and assumption 3.4 in Chen and Pouzo (2012). An unconstrained version of this estimator is used in Breunig (2013) to test for the smoothness of a quantile specification.

I do not provide exact analytical proof for the asymptotic properties of this ad hoc empirical testing procedure. I assess its numerical performance in an exhaustive simulation study in the next section. The results show that the test performs well under a variety of alternative forms and under different degrees of endogeneity and strength of the instrument.

Testing: simulation

To assess the power of the test I simulate two different non-monotone models:

- Model 1 is borrowed from the simulation study of Hall and Heckman (2000). The true function is defined as

  $$m(x) = (15(x-0.5)^3 + M(x-0.5))\mathbb{I}_{[0,0.5]} + M(x-0.5)\mathbb{I}_{(0.5,1]} - \exp\{-250(x-0.25)^2\}\mathbb{I}_{[0,1]},$$

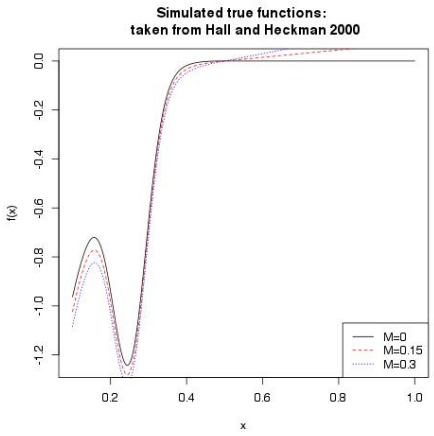
  see figure 2.6a,a). The parameter M controls the amount of "non-monotonicity" with higher values leading to a function closer to a monotone one. I simulate this model for $M = 0.3$.

- Model 2 is borrowed from the study of Bowman, Jones, and van der Gijbels (1998). The true function is defined over the range $(0, 1)$ as

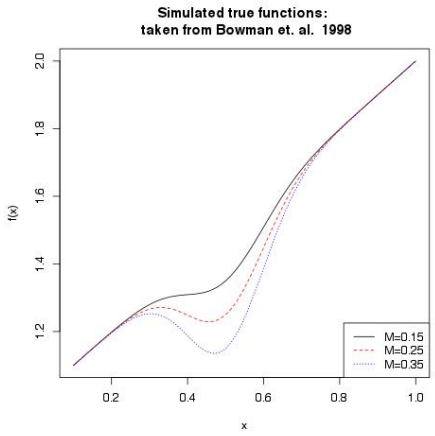
  $$m(x) = 1 + x - a \exp\{-\frac{1}{2}\Big(\frac{x-0.5}{0.1}\Big)^2\},$$

Figure 2.6: Simulated models, Testing

(a) Model 1



(b) Model 2



where the parameter a controls the amount of monotonicity: the function is linear for $a = 0$ and monotone for $a \leq 0.15$, see figure 2.6a,b). I simulate the model for the value $a = 0.35$.

These two models incorporate different types of violations of the monotonicity assumption. Whereas in the first model the interval on which the function is decreasing is small with a big drop, in the second model the drop is moderate but on a longer interval.

I perform the test using the constrained PSMD estimator described in section 2.4.6 with optimal parameters for a nondecreasing specification, that is, model 2 from the last subsection. In both models the power approaches the target level of 5 %. The empirical probabilities to reject the null for $M = 0.3$ and a=0.35 for different cutoffs are shown in tables 2.12 and 2.13.

To assess the size of the test I simulate for each sample size n $N = 100$ Monte Carlo samples of a constant function model. The empirical probabilities to reject the null for a targeted level of 1% are shown below in table 2.14. Thus, the test has both good power and size properties.

Table 2.12: Power Model 1, $M = 0.3$

n	$J_n = 4$	$J_n = 5$	$J_n = 6$	$J_n = 7$	$J_n = 8$	$J_n = 9$	$J_n = 10$
50	0.04						
100	0.09	0.11					
300	0.36	0.31	0.30	0.28			
500	0.65	0.57	0.53	0.50	0.50		
700	0.69	0.67	0.70	0.68	0.68	0.67	
1000	0.93	0.91	0.94	0.91	0.90	0.88	0.86

Notes: Values calculated only for admissible values of the cutoff parameter ($J_n \leq n^{\frac{1}{3}}$)

Table 2.13: Power Model 2, $a = 0.35$

n	$J_n = 4$	$J_n = 5$	$J_n = 6$	$J_n = 7$	$J_n = 8$	$J_n = 9$	$J_n = 10$
50	0.12						
100	0.23	0.23					
300	0.82	0.80	0.78	0.77			
500	0.87	0.92	0.89	0.86	0.84		
700	0.98	0.99	0.98	0.97	0.97	0.98	
1000	1.00	1.00	0.99	0.99	0.97	0.98	0.95

Notes: Values calculated only for admissible values of the cutoff parameter ($J_n \leq n^{\frac{1}{3}}$)

Table 2.14: Size Model 3

n	$J_n = 4$	$J_n = 5$	$J_n = 6$	$J_n = 7$	$J_n = 8$	$J_n = 9$	$J_n = 10$
50	0.01						
100	0.01	0.01					
300	0.01	0.01	0.03	0.02			
500	0.03	0.02	0.02	0.02	0.03		
700	0.00	0.01	0.02	0.04	0.04	0.03	
1000	0.03	0.02	0.01	0.01	0.00	0.01	0.00

Notes: Values calculated only for admissible values of the cutoff parameter ($J_n \leq n^{\frac{1}{3}}$)

2.6 Empirical investigation of the effect of class size on test scores in a Minnesota data set

The study of Cho, Glewwe, and Whitler (2012) evaluates the effect of class size on test scores in public elementary schools in the U.S. state Minnesota. Every year beginning with the school year 1997/98, all 3rd and 5th graders take part in Minnesota Comprehensive Assessment (MCA) test, a standardized test administered by the state of Minnesota. Around 60000 students in each grade are evaluated at the subjects reading and math.[5] The data set contains class sizes and (school-level) average test scores in 8 school years (1997/1998 - 2004/2005). It also contains basic school-level demographic information, such as the percentage students eligible for a subsidized lunch (Free Lunch), the percentage students with limited English proficiency (lep), the percentage of students in a special educational status, and the proportion of different minority groups. The initial number of schools is 1116 for the 3rd graders and 1137 for the 5th graders. However, there is attrition due to various reasons, such as missing demographic information (1-2 %), no publicly available test score data due to too small classes (8-9 %), and test score data for less than two years (7-8 %) (not usable for estimation with school fixed effects). Table 2.15 gives an overview of the attrition reasons. By far the most important attrition reason is missing information on the number of classes in a school. The class size in a school is obtained via dividing the enrollment by the number of classes. Of both variables though, only the enrollment size is publicly available. The number of classes were obtained by the authors of the original study through a telephone survey, however, not for all schools. Table 2.16 contains a summary of descriptive statistics of the available demographic information for both schools with and without class size information. The values for means and standard deviations are very similar. I calculate the p-values for a test for equality of the means of these variables between the two groups of schools. I find no significant differences between the two subsamples. Thus, there is no sign for selection induced by the main reason for attrition.

The final sample contains 2493 3rd grade classes from 484 schools and 2368 5th grade classes from 460 schools. For 3rd and 5th grade, the minimum class size is 14 and 13, the mean class size 23 and 25, and the maximum 34 and 37, respectively. Pictures 2.7a and 2.7b depict the histograms of the corresponding distributions of class size.

2.6.1 A separable econometric model

Denote with X the class size, with Z all other observable variables that influence the school success (Free Lunch, lep and the ethnical composition), and with ϵ all relevant unobservables. Z obtains values in Ω_Z. Since Z is observed, I have to specify a model that incorporates these observed characteristics and define what precisely is the causal effect of class size on test scores. Consider the fully nonparametric model

$$Y = \widetilde{m}(X, Z, \epsilon). \tag{2.6.1}$$

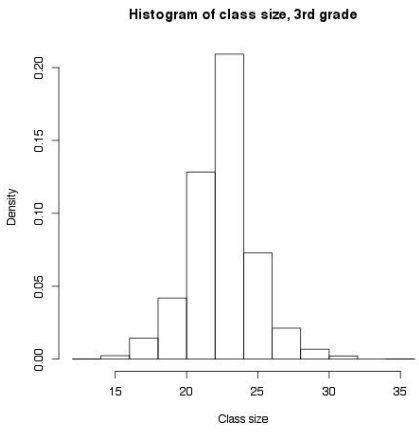
[5] In addition, 5th graders are evaluated in a writing test that consists of 4 parts. However, each of the students is evaluated in only one of these four parts. Therefore, for comparability reasons, this test is excluded from the analysis in the original study and in my paper.

Table 2.15: Sample attrition (Source: Cho, Glewwe, and Whitler (2012))

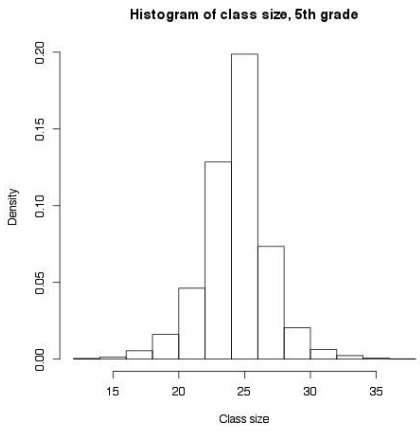
	3rd grade	5th grade
Schools that appear in the data at least once between 1997/98 and 2004/05	1116	1137
Of which: schools missing demographic data for every year	15	32
Schools with < 10 students	88	107
Schools with test participation rate < 90 %	7	7
Schools with <2 years of test score data	80	91
Schools with demographic data, ≥ 10 students, test participation ≥ 90% and with ≥ 2 years of test score data	922	895
Of which: schools with < 2 years class size data	420	413
Schools with demographic data, ≥ 10 students, test participation ≥ 90% and with ≥ 2 years of test score and class size data	502	482
Of which: schools with < 5 years of enrollment data	18	22
Number of schools included in the regression analysis	484	460

Figure 2.7: Histogram class size

(a) 3rd grade



(b) 5th grade



In this general framework, the main ingredient of causal analysis is the first derivative of $\widetilde{m}$ with respect to X, $\partial\widetilde{m}(.,z_0,\epsilon_0)/\partial x$, where z_0 and ϵ_0 are some elements of the domains of Z

and ϵ. In general, $\widehat{m}$ and its derivative cannot be identified without additional restrictions, see Chesher (2003). One way to achieve identification is to impose additive separability of observed and unobserved covariates, see e.g. Darolles, Fan, Florens, and Renault (2011),

$$Y = \overline{m}(X, Z) + \epsilon. \tag{2.6.2}$$

The derivative can be recovered when a valid instrument for the endogenous class size is available. The central object of econometric analysis in 2.6.2 the partial derivative $\partial\overline{m}(., z_0)/\partial x$. Model 2.6.2 allows the effect of class size on test scores to vary with observed characteristics. The empirical test for monotonicity developed in section 2.5 can be easily adapted to that case. It is not clear, however, for which values z_0 the test should be performed. The interpretation of the results will always depend on the value z_0.

To improve the interpretability of the model, I assume additive separability of X and Z which leads to the model

$$Y = m(X) + g(Z) + \epsilon. \tag{2.6.3}$$

In a model with exogenous covariates, the separability assumption is equivalent to imposing homogeneous additive average treatment effects:

$$\mathbb{E}[Y \mid x_1, z] - \mathbb{E}[Y \mid x_2, z] = m(x_1) - m(x_2).$$

We observe that the right hand side does not depend on Z. Assuming homogeneous treatment effects might be at odds with economic theory. An increase of class size might have only a small effect when the parents are able to compensate via private lessons. On the other hand, an increase in class size might have a strong negative effect in schools with students from a predominantly lower social economic background. Thus, there is a tradeoff between the interpretability/complexity of the model and the plausibility of its assumptions.

Nevertheless, 2.6.3 is still a very general model. I do not specify a parametric form for m or g. 2.6.3 contains as a special case the linear model

$$Y = \alpha X + \beta' Z + \epsilon$$

which is assumed throughout the literature. To adapt model 2.6.3 to a panel context, I include school and district random components, so finally I impose the model

$$Y_{i,j,k,t} = m(X_{i,j,k,t}) + g(Z_{i,j,k,t}) + S_j + D_k + \epsilon_{i,j,k,t}, \tag{2.6.4}$$

where $Y_{i,j,k,t}$ is the average test score for grade i in school j in district k in year t, S_j are school and D_k district random effects, and X and Z as above.

Table 2.16: Descriptive statistics for schools with and without class size data (Source: Cho, Glewwe, and Whitler (2012), p-values: own calculations)

	Schools Group 1		Schools Group 2		
	mean	sd	mean	sd	p-value
		Grade 3			
Enrollment	73.3	43.4	75.9	41.3	0.26
Class size	-	-	22.4	4.7	-
Black (%)	6.9	14.4	7.2	14.0	0.70
Hispanic (%)	3.8	7.5	3.8	7.3	1.00
Asian (%)	4.9	9.7	5.0	8.8	0.84
Am. Indian (%)	2.3	7.8	2.1	6.9	0.61
White (%)	81.5	24.2	81.3	23.9	0.88
Male (%)	51.2	7.5	51.1	7.2	0.80
Free Lunch (%)[6]	33.8	22.2	32.3	22.6	0.23
lep (%)	6.0	11.8	6.1	11.1	0.87
Students in special education (%)	11.8	6.5	11.8	5.9	1
Sample size		922		502	
		Grade 5			
Enrollment	80.1	58.3	80.0	45.7	0.97
Class size	-	-	24.4	5.4	-
Black (%)	6.9	13.9	7.4	14.5	0.54
Hispanic (%)	3.4	6.6	3.5	6.8	0.79
Asian (%)	5.1	10.0	5.3	9.8	0.72
Am. Indian (%)	2.3	7.6	2.2	6.4	0.93
White (%)	81.7	24.0	81.1	24.4	0.66
Male (%)	51.3	7.3	51.2	7.2	0.80
Free Lunch (%)	33.1	22.1	32.1	22.9	0.43
lep (%)	5.4	10.8	5.6	10.6	0.74
Students in special education (%)	13.5	6.5	13.4	5.8	0.77
Sample size		895		482	

Notes: Group 1: schools with demographic data, high test participation and ≥ 2 years of test data. Group 2: schools for which there is also data on the number of classes. p-values calculated on the basis of a T-test for equality of means in independent samples.

2.6.2 Empirical instrumental strategy: Hoxby's approach

As discussed in section 2.3, the class size variable X might be endogenous. A key to the econometric analysis is therefore the availability of an instrument. Since the observations of X are on a school level, within-school selection is not of a concern. As a result, an instrumental variable W has to deal successfully only with between-schools selection. Formally, in the context of model 2.6.4, an instrumental variable W has to fulfill the following exclusion restriction (ER):

$$\begin{aligned}&\mathbb{E}[g(Z_{i,j,k,t}) \mid W_{i,j,k,t}] = \mathbb{E}[S_j \mid W_{i,j,k,t}] = \\ &= \mathbb{E}[D_k \mid W_{i,j,k,t}] = \mathbb{E}[\eta_t \mid W_{i,j,k,t}] = 0.\end{aligned}$$

The condition $\mathbb{E}[g(Z_{i,j,k,t}) \mid W_{i,j,k,t}]$ is not necessary for identification. It allows to consider a generalized error term $g(Z_{i,j,k,t}) + S_j + D_k + \epsilon_{i,j,k,t}$ and treat the observables as unobservables. Thus, model 2.6.4 is equivalent to model 2.3.1 and the test for monotonicity can be applied directly. In the rest of the paper, the nonparametric analysis of the functional form of m is performed unconditionally.[7]

I adopt the instrumental variable strategy used by Cho, Glewwe, and Whitler (2012). The idea goes back to the influential study of Hoxby (2000). She follows two different approaches to identify the effect of class size on test score in elementary schools in Connecticut. In the first one, she uses random variation in the enrollment and in the second one a maximum class size rule as an instrument for class size. Since in Minnesota only few districts impose a class-size caps, only the first approach can be applied to this dataset. The idea for constructing an instrument can be summarized as follows. The enrollment for each year can be split into two parts: one part that depends on parents, students and community characteristics and changes slowly and continuously over time, and a random variation part. Following the notation of Cho, Glewwe, and Whitler (2012), denote the deterministic part of enrollment in grade i of school j in district k in year t with $E_{D,i,j,k,t} = E_{D,i,j,k,t}(X_{i,j,k,t}, v_{i,j,k,t})$, where $X_{i,j,k,t}$ and $v_{i,j,k,t}$ are observed and unobserved parents, students and community characteristics. $E_{D,i,j,k,t}$ is modeled as a continuous function of time,

$$E_{D,i,j,k,t} = g(t) \tag{2.6.5}$$

which can be approximated by a polynomial, $\log(g(t)) \approx \alpha_{0,i,j,k} + \alpha_{1,i,j,k}t + \cdots + \alpha_{p,i,j,k}t^p$.[8] The total enrollment is modeled as

$$E_{i,j,k,t} = E_{D,i,j,k,t}U_{i,j,k,t}, \tag{2.6.6}$$

where $U_{i,j,k,t}$ denotes a random variation component. Taking 2.6.5 and 2.6.6 together yields the regression equation

$$\log(E_{i,j,k,t}) = \alpha_{0,i,j,k} + \alpha_{1,i,j,k}t + \cdots + \alpha_{p,i,j,k}t^p + \log(U_{i,j,k,t}). \tag{2.6.7}$$

Under the assumption that the number of classes is fixed over time and that the polynomial specification is correct, $\log(U_{i,j,k,t})$ is a valid exclusion restriction as it is independent of the

[7] Nevertheless, I use the observed covariates Z to analyze the validity of the exclusion restriction.

[8] Cho, Glewwe, and Whitler (2012) approximate $g(t)$ choosing $p = 3$ for their benchmark estimations and $p = 4$ for the robustness checks.

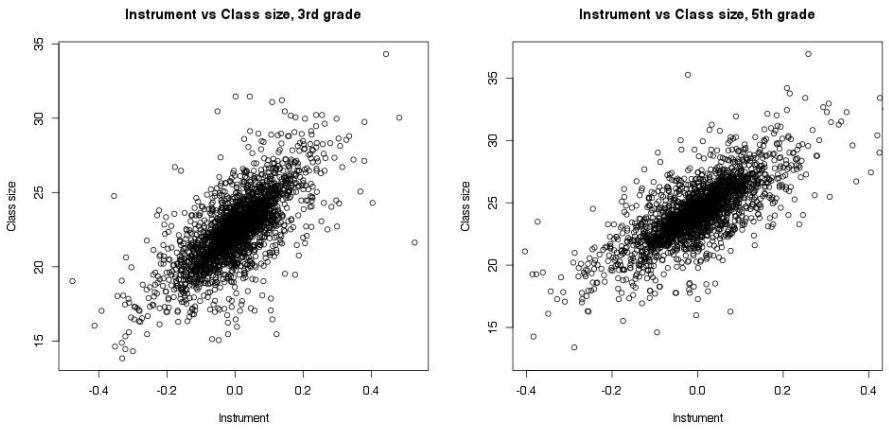
unobserved characteristics $v_{i,j,k,t}$. The approach of Cho, Glewwe, and Whitler (2012) and Hoxby (2000) is to estimate equation 2.6.7 with OLS and to use the OLS residuals estimates $\widehat{\log(U_{i,j,k,t})}$ as instruments for the endogenous class size.[9] Cho, Glewwe, and Whitler (2012) use 17 years of enrollment data (1988/89-2004/05) to obtain estimates for $\log(U_{i,j,k,t})$.

Analysis of the instrumental strategy A good instrument has two properties: it explains a considerable part of the variation of the endogenous regressor and it is a valid exclusion restriction. In this paper, I suggest a new empirical approach for the analysis of these properties.

First, the correlation of the instrument with the class size is 0.71 for third graders and 0.69 for 5th graders, indicating that the instrument is strong. Figures 2.8a - 2.8b show plots of the class size against the instrument. Higher (lower) values of random enrollment variations lead to higher (lower) class sizes, which leads to the strong positive correlation.

Figure 2.8: Correlation between instrument and class size

(a) 3rd grade (b) 5th grade



The exclusion restriction is not testable, but indirect evidence for its plausibility can be provided. The starting point for my analysis are the possible reasons for a violation of the exclusion restriction. All three variables - the instrument, the (average) class size and the (average) test scores are school-level variables and therefore within-school selection is not of a concern. On a between-schools level, there are two possible reasons for a violation of the exclusion restriction. First, the change of parents, students, and community characteristics might be discontinuous. In that case, the approximation with a continuous function 2.6.7 would not be proper. Discontinuities might arise because of a structural break in the time trend of these characteristics. Possible reasons are big layoffs in the district of the school, migration waves, unusually large cohorts, or other micro- and macro-level shocks. On the contrary, Tiebout sorting and other related endogeneity sources do not violate the exclusion restriction, as long as the patterns of sorting vary smoothly with the district characteristics over time. Second, a polynomial of third degree might be a poor approximation of the

[9] Below, I briefly discuss implications of using an estimated instrument.

underlying continuous evolvement of parents, students, and community characteristics. As a consequence, the estimated random part might capture some of the influence of these characteristics.

A first strategy to obtain evidence for (or against) the plausibility of the exclusion restriction is to verify the possible reasons for violations of the exlcusion restriction (an indirect strategy). One way to address the first problem is to check for structural changes over time in observed characteristics. Tests for structural breaks in panel data context are developed in Han and Park (1989) and Kao, Trapani, and Urga (2007). The identification assumption of this strategy is that observed and unobserved characteristics behave similarly. Thus, the absence of structural changes in observed characteristics provides indirect evidence that also unobserved characteristics evolve continuously. I do not have data on observed characteristics for the 17 years used to calculate the instrument and cannot test for structural break.

One way to address the second problem would be to perform robustness checks with polynomials from higher degrees. If the coefficients of the higher order terms are small and insignificant, and if the results from the estimation of the causal effect of class size remain unchanged under the new specifications, then one could conclude that the initial polynomial specification is nearly correct. The identifying assumption here is that there are no structural breaks in observed and unobserved characteristics, i.e. that the first type of violation of the exclusion restriction is not existent. For the same reason as above, I cannot perform this type of exercise (the robustness checks of Cho, Glewwe, and Whitler (2012) in the context of a linear model reveal stable results).

Instead of checking for a violation of these two sufficient conditions for the validity of the exclusion restriction (smooth pattern over time and correct polynomial specification), I address possible violation of the exclusion restriction itself. My main identification assumption is that observed and unobserved characteristics behave similarly in the sense, that a violation of $(Z, \epsilon) \perp\!\!\!\perp W$ can be verified by checking whether $Z \perp\!\!\!\perp W$ holds.

I perform a test for independence of observed covariates and the instrument (pairwise). The test based on a nonparametric metric entropy and is described in Li, Maasoumi, and Racine (2009) Racine. The test statistic is defined as

$$S_\rho := \frac{1}{2} \int_{\Omega_Z} \int_{\Omega_W} ((\widehat{f}_{Z,W})^{\frac{1}{2}} - (\widehat{f}_Z)^{\frac{1}{2}} (\widehat{f}_W)^{\frac{1}{2}})^2 dz dw, \tag{2.6.8}$$

where $\widehat{f}$ is a kernel estimator of f. Big values of the test statistic indicate deviation from independence. The distribution of the test statistic is obtained via bootstrapping. Table 2.17 contains the p-values for the (pairwise) independence of the instrument and the observed covariates lep, Free Lunch and White (the proportion of white students in the school) obtained with 100 bootstrap replications. The p-values are very large and the null hypothesis (independence) cannot be rejected.

Observe, that independence of the instrument and ϵ is not a necessary condition for W to be a valid instrument. W only has to satisfy the much weaker conditional mean independence assumption, $\mathbb{E}[\epsilon \mid W] = 0$. Under the same identification assumption as above, this assumption can be verified by looking at $\mathbb{E}[Z \mid W] = 0$. In particular,

$$\mathbb{E}[Z \mid W] = \mathbb{E}[Z] \tag{2.6.9}$$

would provide indirect evidence that $\mathbb{E}[\epsilon \mid W] = 0$. Since both Z and W are observable, 2.6.9 is a testable hypothesis. I estimate left hand side of 2.6.9 with a Nadaraya-Watson (NW)

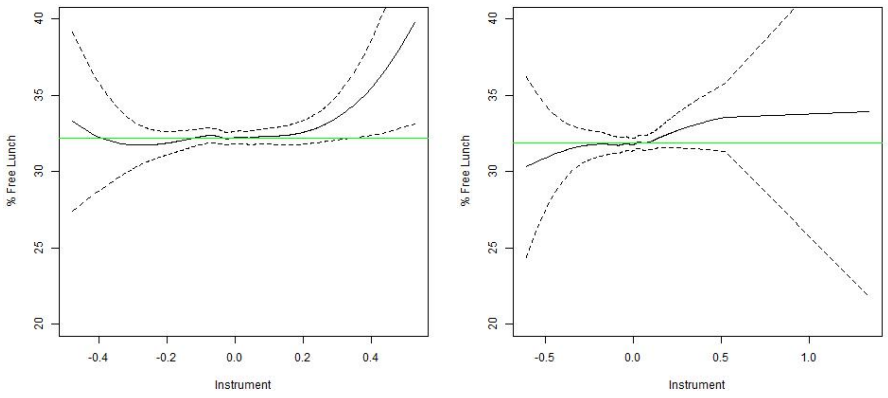
Table 2.17: Testing for independence of instrument and observed characteristics. p-values obtained with 100 bootstrap replications)

	3rd grade	5th grade
lep	0.67	0.70
Free Lunch	0.85	0.74
White	0.49	0.53

kernel estimator and right hand side with a sample mean. The resulting estimates are shown in figures 2.9a-2.11b. The green thick line is the unconditional estimator of the observed covariate. The dashed lines are the 95% confidence bounds of the NW estimator. In all 6 cases, the unconditional mean lies between the confidence bounds almost on the whole range. This finding provides a substantial evidence that the null hypothesis 2.6.9 is plausible, and by means of the identification assumption, that $\mathbb{E}[Z \mid W] = 0$ holds.

Figure 2.9: Conditional expectation of Free Lunch

(a) 3rd grade (b) 5th grade



Remark Note that one possibility to reduce the likelihood for a violation of the exclusion restriction, as suggested by Cho, Glewwe, and Whitler (2012) and Hoxby (2000), is to model district-level enrollment in equation 2.6.7 instead of school-level enrollment. This strategy would account for an usually large school enrollment that could otherwise lead to a structural break if a larger proportion of parents send their children to another school in the district. Such large fluctuations are also possible on a district level. However, sending the children to a school in a different district is associated with higher costs and is therefore less likely.[10] One

[10]Nevertheless, to account for this possibility, Cho, Glewwe, and Whitler (2012) use number of births in

Figure 2.10: Conditional expectation of LEP

(a) 3rd grade (b) 5th grade

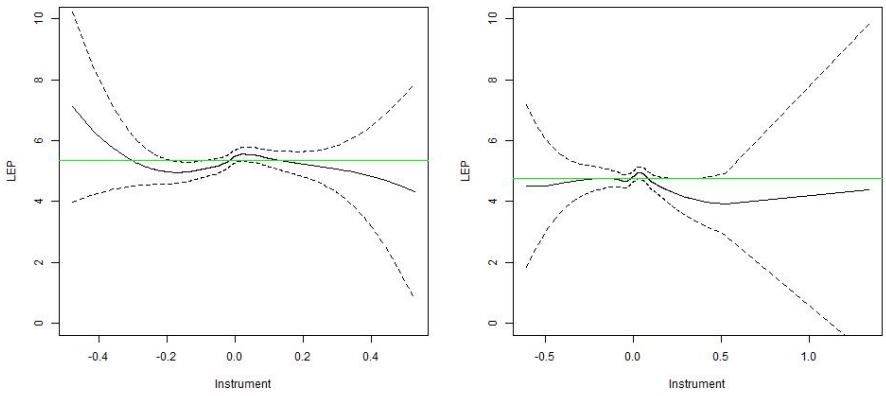
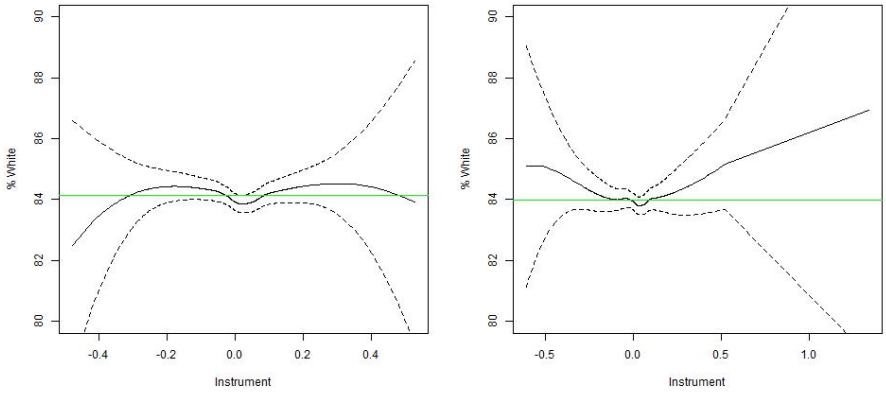


Figure 2.11: Conditional expectation of White

(a) 3rd grade (b) 5th grade



drawback associated with the district-level enrollment as an instrument is that it explains a smaller part of the variation in the class size, that is, it is a weak instrument. I have no data on the district level and cannot pursue this strategy.

Minnesota cities 8 or 10 years prior to the year of test as an instrument (for 3rd and 5th graders, respectively). This strategy leads only to a very weak instrument, as indicated by the authors of the original study. Therefore, I will not pursue it further in my paper.

2.6.3 Main results: shape analysis of the effect of class size on test scores

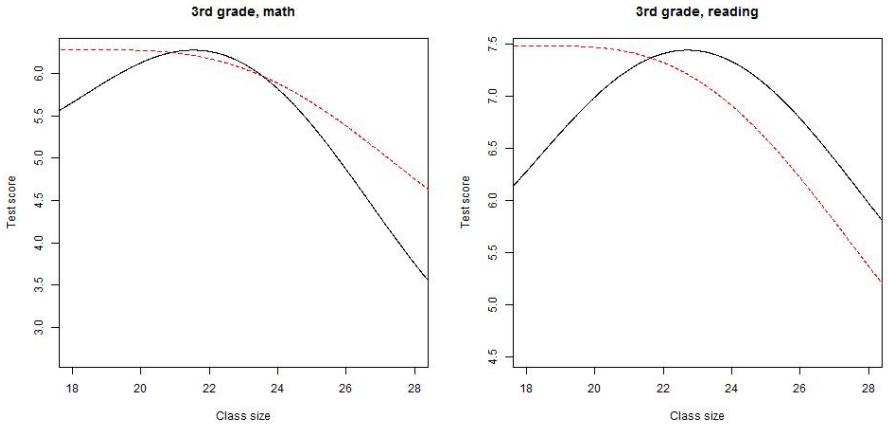
After analysing the plausibility of the exclusion restriction, I now present my main results. I perform shape analysis of the effect of class size on test scores in three different stages. First, I plot nonparametric unconstrained and constrained estimates of the regression function. The graphical analysis delivers first key insights about the shape of the causal effect. Second, to legalize these findings, I test the significance of the result performing the empirical test for monotonicity developed in section 2.5. Finally, I demonstrate that the causal effect can be approximated with a polynomial of second degree.

Although I use all observations for the estimations and testing procedures, most of the graphical analysis (including the search for a parametric specification) is done on the restricted ranges $[18, 28]$ for third grade and $[20, 30]$ for 5th grade. Those two ranges contain 95% and 94% of all points respectively, see the histograms of class size on figures 2.7a and 2.7b. The reason is that at the boundary the nonparametric estimator does not perform well. For the purposes of the graphical analysis, these ranges are sufficient. In a typical policy analysis, most decisions are made precisely in those ranges. I will refer to them as "the middle range".

Graphical analysis I first estimate the regression function nonparametrically using the unconstrained penalized sieves estimator 2.4.15 and its constrained nonincreasing version 2.4.16- for different choices of the cutoff parameter J. The estimates for the cutoff parameter 7 and $\alpha_n = 0.001$ for the middle range are depicted in figures 2.12a-2.13b for both math and reading.

Figure 2.12: Unconstrained and constrained estimates, 3rd grade

(a) math (b) reading

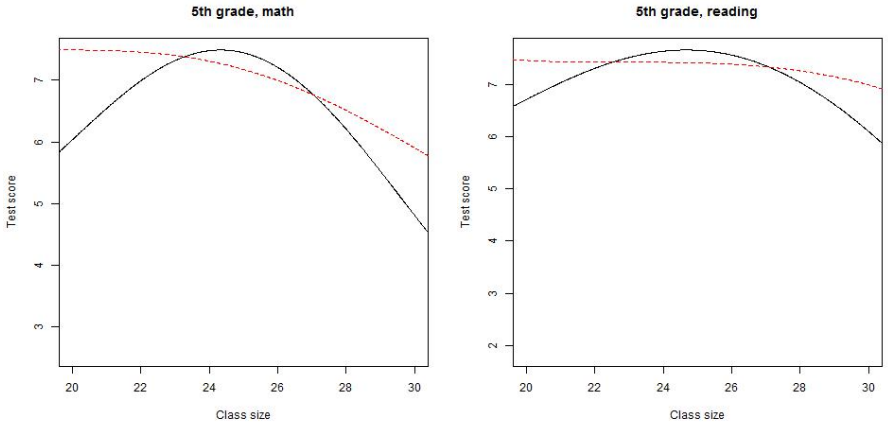


The x-axis represents the class size and the y-axis the test scores. The black thick lines represent the unconstrained estimates, the red dashed lines their constrained counterpart. The unconstrained estimates are nonmonotone and concave. At first they are increasing, indicating a positive effect of class size on test scores, and subsequently decreasing.[11]

[11] As noted, the grades are normalized by the standard deviation (therefore they cannot be used to predict average score for a given average class size).

Figure 2.13: Unconstrained and constrained estimates, 5th grade

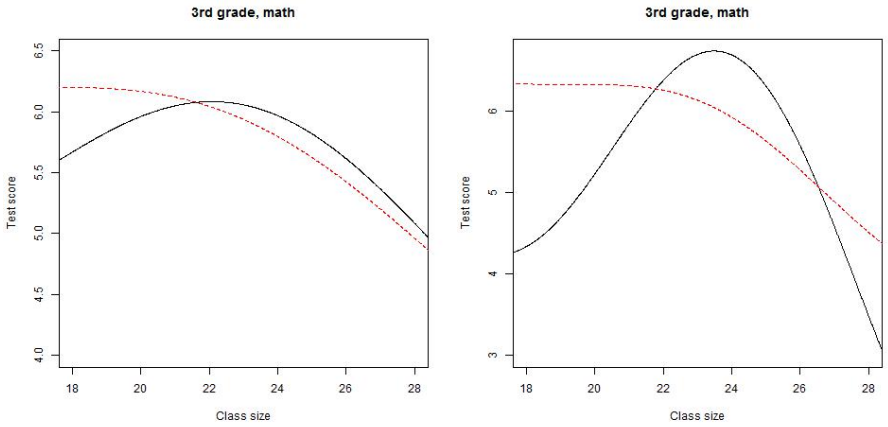
(a) math (b) reading



Although the simulation delivered a clear pattern of how to set the penalty constant, the choice for the cutoff parameter according to the sample size was not perfectly clear (although generally also rising with n). The unconstrained nonparametric estimates depend upon the choice of the cutoff parameter in the following way. Increasing the cutoff parameter (the number of basis functions used for the Least Squares estimate) for a fixed sample size makes the estimate more "nonmonotone". Figures 2.14a and 2.14b show the estimated regression function for 3rd grade math the cutoff parameter equal to 5 and 9. The pattern of the nonmonotonicity remains stable.

Figure 2.14: Unconstrained and constrained estimates, 3rd grade math

(a) M=5 (b) M=9



Remark (sample splitting) An alternative ad hoc way to check graphically for possible

nonmonotone effects would be the following method. First split the sample into two parts: one sample with all observations with class size below a threshold $\bar{c}$, e.g. $\bar{x} = 25$, and one sample with all observations with class size above the threshold. Then in those two samples estimate a linear model and compare the slopes. A difference in the signs of the slopes could be interpreted as a nonmonotonicity. The advantage of this approach is its simplicity. It has three major disadvantages though. First, the result might depend sensitively on the chosen threshold. Second, as an average approach it is not appropriate for analysing the overall functional form of the regression, but only for detecting nonmonotonicity. And third and most important, the exclusion restriction in each subsample breaks down due to the condition $x \leq \bar{x}$ ($x > \bar{x}$). To see this, suppose that all random variables in the model are defined on some probability space $(\Omega, \mathscr{A}, P)$ with P being the measure and $\mathscr{A}$ a sigma field. Denote with $\sigma_1 := \sigma(1\{X \leq \bar{x}\})$ and $\sigma_2 := \sigma(1\{X \leq \bar{x}\}, W)$ the sigma fields generated by the random variables $1\{X \leq \bar{x}\}$ and $(1\{X \leq \bar{x}\}, W)$, respectively. Then $\sigma_1 \subset \sigma_2 \subset \mathscr{A}$ and $\mathbb{E}[\epsilon \mid X \leq \bar{x}, W] = \mathbb{E}[\epsilon \mid \sigma_2]$. With the smoothing law of iterated expectations,

$$\mathbb{E}[\mathbb{E}[\epsilon \mid \sigma_2] \mid \sigma_1] = \mathbb{E}[\epsilon \mid \sigma_1],$$

or equivalently

$$\mathbb{E}[\mathbb{E}[\epsilon \mid X \leq \bar{x}, W] \mid X \leq \bar{x}] = \mathbb{E}[\epsilon \mid X \leq \bar{x}].$$

If $\mathbb{E}[\epsilon \mid X \leq \bar{x}, W] = 0$ a.s., then it must hold $\mathbb{E}[\epsilon \mid X \leq \bar{x}] = 0$ a.s. which in general is a contradiction to the endogeneity of X. Therefore, $\mathbb{E}[\epsilon \mid X \leq \bar{x}, W] \neq 0$ and the exclusion restriction in the subsamples is violated (note, that this conclusion is not true if the threshold $\bar{x}$ is chosen such that $P\{X \leq \bar{x}\} = 1$). Intuitively, since ϵ and X are not independent, conditioning on X will create endogeneity and potentially create a bias in the estimates.

Testing for monotonicity The (unconstrained) graphical nonparametric analysis reveals a nonmonotonic pattern of the regression function that is common for all grades and subjects in the middle range of class sizes. In order to assess whether this pattern is significant, I test for the monotonicity of causal effect using the procedure developed in 2.5. The null hypothesis is defined as

$$H_0 : m' \leq 0.^{12} \tag{2.6.10}$$

Table 2.18 summarizes the results. The first column of the table contains the value of the smoothing parameter of the test statistic M. It specifies the number of basis functions used to calculate the test statistic. It is allowed to grow with the speed $o(n^{\frac{1}{3}})$ to ensure that the null distribution is a standard normal distribution. The columns with header n contain the sample size. The constrained estimate used for the test has been calculated with a cutoff parameter equal to 7 and a penalty constant $\alpha_n = 0.001$. p-values that are smaller than 0.00001 are reported as 0, which is here the case for all grades and subjects and test specifications ($m = 4$ to $m = 9$). Therefore the test rejects the monotonicity hypothesis.

This is a novel result. If the exclusion restriction is valid, this finding has two important implications.[13] First, the test rejects the special case $m_0 = 0$. That is, the class size must have a

[12] I also specify the opposite as a null, namely that the function is monotonically increasing, the results are similar as those in the first case and are not discussed separately.

[13] Note, that the instrument has been estimated, which potentially creates endogeneity. However, the estimate converges to its population value with a parametric rate. Therefore, provided that the population value is a

Table 2.18: p-Values, test for monotonicity, dataset of Cho et al (2012)

Whole Sample								
	3rd grade				5th grade			
	math		reading		math		reading	
M	n	p-value	n	p-value	n	p-value	n	p-value
4 -9	2493	0	2493	0	2368	0	2390	0

significant effect on the average test scores at least on a segment of the class size range. This conclusion is at odds with the findings of Hoxby (2000), who uses the same instrumental variable strategy but with a different data set and with a parametric specification. Second, the test rejects the linear model which is the most common specification in the empirical literature. Therefore, a more flexible way to model the causal effect is needed. This conclusion differs from the negative and significant results of Cho, Glewwe, and Whitler (2012)

Parametric specifications One important question for the applied research is whether there exists a parametric specification that is a good approximation of the causal effect. Parametric specifications are straightforward to implement and are easier to interpret in terms of causal analysis. Moreover, they are less data intensive than their nonparametric counterparts when there are multiple covariates. All these advantages make parametric specifications attractive tools for policy analysis.

If we assume that the causal effect is continuous, then polynomials are a good starting point for the search of a parametric specification. First, one can approximate continuous functions arbitrary well with polynomials. Second, polynomials with degree higher than 2 can generate a nonmonotone regression function, which is a necessary property in view of the preceding graphical and testing results. I will focus on the middle range of the class sizes, that is on $[18, 28]$ for third grade and $[20, 30]$ for fifth grade. The form of the unconstrained estimator in figures 2.12a- 2.14b indicates that a polynomial of second degree is a potential candidate for a parametric specification (in the middle range). Second degree approximations of the causal effect are shown in figures 2.15a-2.16b.

The black continuous lines are as before the unconstrained nonparametric estimates. The blue dashed lines are obtained via approximating the IV estimates with a second degree polynomial (estimating a second degree polynomial directly with the data only in the middle range creates potential endogeneity for precisely the same reason as in the argument with the sample splitting). The fitted values are obtained by regressing the IV estimates on a polynomial using ordinary OLS. As the figures reveal, the parametric specification works surprisingly well.

To exclude the possibility that these findings are only a matter of the concrete sample, I perform a test for a parametric specification. Typically, the decision how many polyno-

valid exclusion restriction, this does not pose a problem for the asymptotic distribution of the test statistic. I am thankful to Gerard Van den Berg for pointing this out to me.

Figure 2.15: Unconstrained and polynomial estimates, 3rd grade

(a) math (b) reading

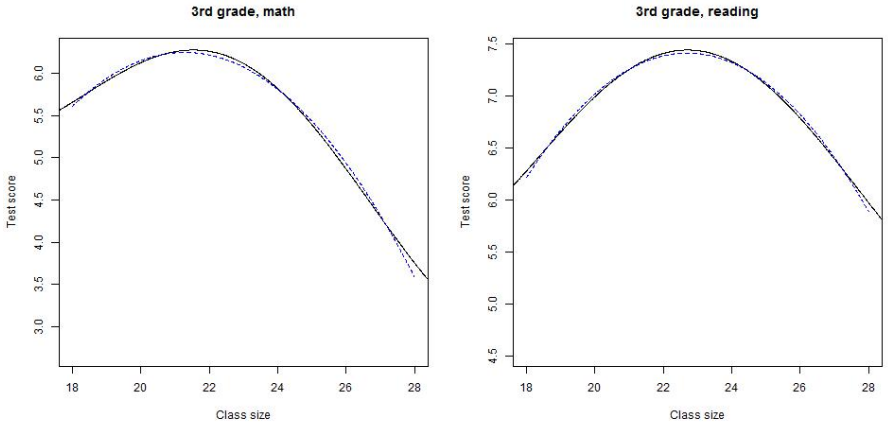
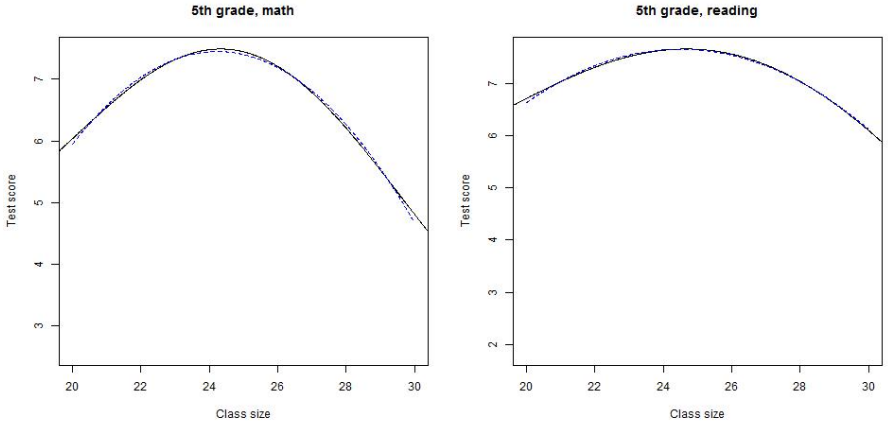


Figure 2.16: Unconstrained and polynomial estimates, 5th grade

(a) math (b) reading



mial terms to include in the regression is based on the significance of the estimates ot the coefficients. For example, if the estimate of β_4 in the model $Y = \beta_0 + \beta_1 c + \beta_2 c^2 + \beta_3 c^3 + \epsilon$ is insignificant and the estimates of β_0, β_1, and β_2 are significant, then c^3 is omitted and the causal relationship is modeled as a polynomial of second degree, see for example Akerhielm (1995) in the context of splines. The underlying assumption is that the true regression function belongs to a subclass of the class of polynomials of degree $k \leq 3$. The null distributions of the estimators $\widehat{\beta}_0, \ldots, \widehat{\beta}_3$ (that is, the distribution of $\widehat{\beta}_i$ when $\beta_i = 0$) are valid only under this assumption. Although this is a useful check, the motivation of this study is more agnostic. I do not assume linearity in the coefficients. Instead, I test the parametric specification against

a nonparametric alternative, that is, I test the whole model. The null hypothesis is

$$H_p: \quad \text{there exists a polynomial } p \text{ of degree} \leq k, \text{ such that} \quad Y = p(c) + \epsilon. \tag{2.6.11}$$

The alternative is that there is no such polynomial. Tests for parametric specifications are developed in Horowitz (2006) and Breunig (2012). In both papers, the test is based on an empirical distance between a parametric IV estimator (the specification under the null) and the data. I implement the series-based estimator by Breunig (2012). I restrict the analysis to $k = 2$ and $k = 3$, that is, to quadratic and cubic polynomials. Under the null, I estimate the regression function m_0 with the 2SLS method using W, W^2, and W^3 as instruments for X, X^2, and X^3, respectively. The test statistic has the standard normal distribution under the null. The results are shown in table 2.19. As above, the first column contains the values of the cutoff parameter of the test statistic. The quadratic specification is rejected for the third grade reading data and the cubic specification for the 5th grade math, both for the whole sample and conditionally on LEP. All other specifications have p-values higher than 10%, although the p-values of the quadratic specification for 5th grade reading are close to 10% (and for m=6 it is actually below). Based on these results, both quadratic and cubic specifications are eligible candidates for the functional form describing the causal effect of class size on test scores.

Table 2.19: p-Values, Test for Parametric Specification, dataset of Cho et al (2012)

Whole Sample											
	3rd grade math			3rd grade reading			5th grade math			5th grade reading	
m	n	d=2	d=3	n	d=2	d=3	n	d=2	d=3	n	d=2	d=3
4	2493	0.18	0.19	2493	0.09	0.16	2368	0.24	0.26	2390	0.14	0.14
5	2493	0.33	0.16	2493	0.08	0.13	2368	0.26	0.18	2390	0.13	0.46
6	2493	0.44	0.41	2493	0.063	0.29	2368	0.19	0.21	2390	0.09	0.35
7	2493	0.69	0.40	2493	0.051	0.39	2368	0.24	0.001	2390	0.13	0.91
8	2493	0.95	0.82	2493	0.04	0.52	2368	0.20	0	2390	0.15	0.65
9	2493	0.64	0.93	2493	0.04	0.71	2368	0.15	0	2390	0.11	0.45
% Limited english proficiency ≤ median of % LEP											
	3rd grade math			3rd grade reading			5th grade math			5th grade reading	
m	n	d=2	d=3	n	d=2	d=3	n	d=2	d=3	n	d=2	d=3
4	1815	0.26	0.27	1813	0.11	0.20	1756	0.30	0.27	1783	0.14	0.14
5	1815	0.73	0.97	1813	0.1	0.17	1756	0.37	0.20	1783	0.16	0.37
6	1815	0.85	0.92	1813	0.1	0.51	1756	0.28	0.26	1783	0.12	0.27
7	1815	0.13	0.21	1813	0.07	0.66	1756	0.33	0.002	1783	0.17	0.69
8	1815	0.21	0.28	1813	0.06	0.99	1756	0.33	0	1783	0.28	0.82
9	1815	0.07	0.15	1813	0.04	0.25	1756	0.27	0	1783	0.23	0.73
% Subsidized lunch ≤ median of % Subsidized lunch											
	3rd grade math			3rd grade reading			5th grade math			5th grade reading	
m	n	d=2	d=3	n	d=2	d=3	n	d=2	d=3	n	d=2	d=3
4	1390	0.42	0.45	1402	0.15	0.12	1330	0.87	0.77	1351	0.29	0.36
5	1390	0.73	0.64	1402	0.11	0.11	1330	0.48	0.67	1351	0.61	0.46
6	1390	0.86	0.97	1402	0.16	0.12	1330	0.50	0.23	1351	0.49	0.86
7	1390	0.96	0.92	1402	0.27	0.25	1330	0.09	0.05	1351	0.45	0.92
8	1390	0.69	0.74	1402	0.23	0.22	1330	0.16	0.02	1351	0.36	0.90
9	1390	0.82	0.83	1402	0.34	0.58	1330	0.02	0.001	1351	0.66	0.43

2.6.4 Nonmonotone class size effect: theoretical background and implications for empirical research

The empirical evidence from the last section is at odds conventional economic theory on the educational production function. Typically, economic theories predict a negative class size effect. They provide two main channels for this effect. First, an increasing size leads to a decrease in the teacher's attention and learning support that each of the student receives, see Correa (1993) for an economic model and Blatchford, Bassett, and Goldstein (2003) for an empirical study. Second, when the class size becomes considerably large, the discipline in the class during the lessonbecomes worse and there are more often lessons disruptions which leads to negative external effects, see Lazear (2001). My empirical results revealed, however, that the regression function is nonmonotone in the middle range: it increases for class sizes 18-25 and decreases for class sizes 25-32. Thus, the increasing part cannot be explained with conventional economic theory.

A possible solution to this problem might found in social cognitive learning theories. Theories that that explicitly model peer effects generally predict a positive relationship between class size and school success, Schunk (1991), Schneeweis and Winter-Ebmer (2007) and Sacerdote (2011). An increasing class size increases the individual probability for interaction with peers which has a positive influence on the cognitive and noncognitive abilities. In addition, in bigger classes there are c. p. more good learners. While poor students profit from their highly skilled peers, good students seem not to be negatively influenced by the presence of low achievers, Schneeweis and Winter-Ebmer (2007).

Taking these considerations into account, an integrated theory might be able to explain nonmontone causal effects. In this subsection, I brieflty discuss possible models of educational production functions that 1) combine effects predicted by conventional economic theory and social cognitive learning theories, and 2) generate a nonmonote overall effect. In particular, I present a simplified production function that incorporates

- the negative effects of an increasing class size resulting from decreasing attention of the teacher that each student receives,
- the negative effect of an increasing class resulting from decreasing quality of the discipline during the lessons, and
- the positive effect of an increasing class size resulting from an increased intensity of social interaction.

Any study that analyzes the effects of those (or others) separate channels is faced with two difficulties. First, data sets rarely contain information about social interactions, teacher's behavior and disruptions of the lesson. Second, small classes (in the range 1-15) are almost never observed, especially in public schools. To build a model of an education production function that reflects my nonmonotone empirical findings, I follow the intuition incorporated in standard economic models and consider the findings in the empirical educational and peer effects studies.

First, it is plausible to assume that the importance of the teacher's attention is highest precisely for small class sizes. While the teaching style in small classes resembles private tutoring, it is likely to be focused on group related activities in bigger classes, see Correa

(1993) for a model that predicts a switch in the teachers' behavior. Therefore the loss of personal learning support or tutoring from one additional student is likely to be highest for smaller classes and decrease when the class size increases. Figure 2.17 depicts such a relationship with individual school success due to teachers attention $a_i(x)$ modeled as $a_i(x) = c_1/x1\{x \leq \bar{x}\} + \bar{x}1\{x > \bar{x}\}$, where c_1 is a constant and $\bar{x}$ is a threshold class size.

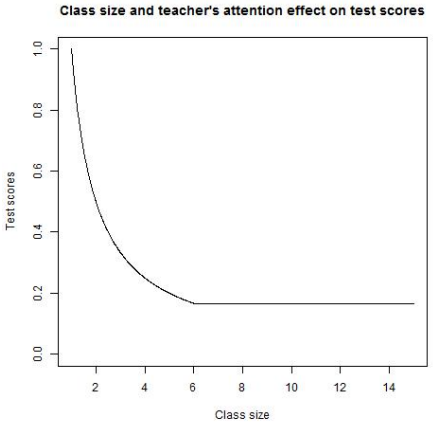


Figure 2.17: The effect of the attention of the teacher on school success

With rising class size the additional school success due to personal mentoring from the teacher decreases up a threshold. After this threshold, the teacher focuses only on group activities, so the level of input for the students stays constant. This relation is also is supported by evidence from the empirical study of Blatchford, Bassett, and Goldstein (2003), who measures in a unique panel study of the effect of class size on test scores the number of interactions between a teacher and the students as well as the discipline in a class.

Second, it is likely that disruptions of the class due to bad discipline become particularly relevant when the class size is very high and are of a less concern for smaller class sizes. Such considerations can be found in for example Lazear (2001), and are supported by the empirical findings of Blatchford, Bassett, and Goldstein (2003). It might therefore be plausible to model the negative discipline effects as a constant up to a threshold class size and as a decreasing function from the threshold on. Such an effect could be generated by the function $d(x) = d_l + 1\{x > d_l\}g(x)$, where d_l is some threshold class size and $g(x)$ is an increasing continuous function with $g(d_l) = 0$, such as $g(x) = x^2 - d_l^2$.

Taking my empirical findings as a starting point, at least on the first part of the middle range the positive effect should be dominating the other two. In addition, with the above considerations, the (negative) discipline effect should be modeled as the dominating one for big class sizes. For smaller class sizes, one has to be agnostic. If we allow the effect from decreasing teacher's attention to dominate the positive effects from interaction, then the regression function will be decreasing for small class sizes, increasing on a segment in the middle range, and then again decreasing. Such a pattern can be generated for example by polynomials from third degree, splines, or other flexible functional forms (using splines or polynomials would also reflect my finding that the causal effect in the middle range is well approximated by a polynomial of second degree). I now discuss briefly two different

possibilities.

Let y be a measure for individual school success, f be the education production function, x the class size and z all other factors that influence y, such as ethnic composition, school resources, teacher quality and others. For simplicity, I abstract from interactions of x and z by fixing z at some level z_0. Then a polynomial specification of third degree for $f(., z_0) = f_0(.)$ is

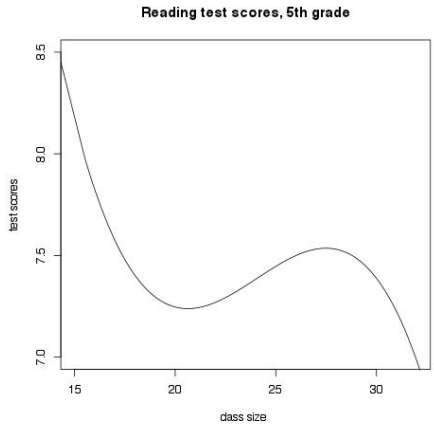
$$f_0(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3,$$

where the coefficients α_i might depend on the point z_0. Figures 2.18a - 2.18b depict the cubic 2SLS estimates for 5th graders and both subjects in the Minnesota dataset.

Figure 2.18: Cubic polynomials, estimates for 5th grade

(a) math

(b) reading



Sufficient conditions on α_i for f_0 to generate the pattern described above (and depicted on 2.18a - 2.18b) are $\alpha_3 < 0, \alpha^2 - 3\alpha_1\alpha_3 > 0$. This model has the disadvantage that the coefficients are not directly interpretable. One possibility would be to require additionally α_0, α_2 to be positive, α_1 to be negative, and to interpret i) α_0 as the test score of a student in a one-student class, ii) α_1 as the effect of one additional student on y resulting from decreasing teacher's attention, iii) α_2 as the social interactions effect and iii) α_3 as the discipline effect. By fixing the coefficients at appropriate values, this would give the dominance described above. Nevertheless, this interpretation would contradict the nonlinearity of the teacher's attention effect and the considerations about the form of the two other effects.

One other possibility for a flexible functional form of f_0 would be to model the three effects separately and impose additive separability. Using the notation introduced above, the overall production function could be defined as $f_0(x) = a(x) - d(x) + i(x)$, where the interaction effects with peers are modeled to be proportional to class size, $i(x) = c_3 x$. This function is continuous and by proper choice of c_1, c_3, d_l and $g(x)$ it can generate the pattern depicted above. If for example $\bar{x} < d_l$, then by assuming $c_3 < -\partial/d(x)\partial x$ on $[d_l, \infty]$ and $\partial a/\partial x + c_3 < 0$ on $[0, \bar{x}]$, one becomes a function decreasing on $[0, \bar{x}]$, rising on $[\bar{x}, d_l]$ and decreasing on the rest of the positive real line.

Several studies have explicitly or implicitly acknowledged the possibility for a non-monotone causal effect. The lack of theoretical interest or own empirical support was probably the reason why these studies did not elaborate on that matter. For example, Lazear (2001) acknowledges that higher class size potentially yields positive peer effects. He argues, however, that since increasing the class size decreases the costs, if students pay the value of the education they receive, adding of extra students will generally take place for class sizes where the negative effects dominate the positive ones. Thus, the possible non-monotone relationship is not of interest, because the economic optimization process has ruled out that decisions are made in the range of class sizes where non-monotonicities occur. As a further example, Dobbelsteen, Levin, and Oosterbeek (2002) acknowledge that the effect of an increase in class size consists of two, opposite in their directions, components. First, with increasing class size, each of the students receives less attention from the teacher. Second, the probability to find a similar peer increases. The latter factor leads to an enhanced learning and thus to a positive effect on the cognitive and noncognitive abilities. The empirical evidence of Dobbelsteen, Levin, and Oosterbeek (2002), however, only supports the dominance of the second factor.

Estimating a linear model when the causal effect is nonmonotone represents a model misspecification. The sign and significance of the class size coefficient may depend sensitively on the form of the regression function. Moreover, differences in the observed range of class sizes between data sets may lead to substantial differences in the estimates, even if the regression function does not differ by much. Model misspecification is a potential explanation of the class size controversy as the following simplified simulated example demonstrates. Suppose that the true regression function is defined as

$$m_0(x) = -4x^3 + 21x^2 - 18x - 12, \tag{2.6.12}$$

see figure 2.19.[14] Suppose also that there are two different samples of observations on test

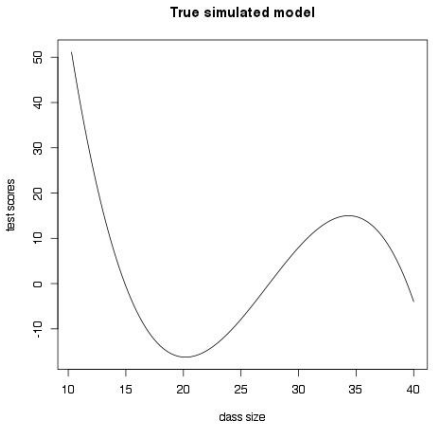


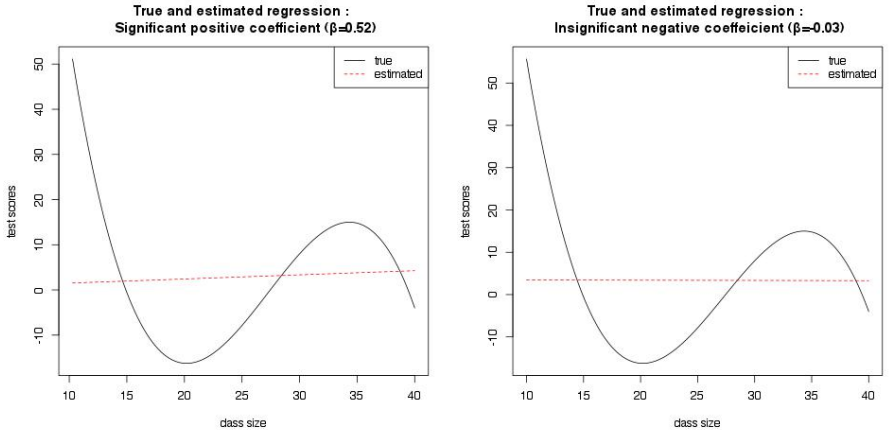
Figure 2.19: Simulated true regression

[14]The independent variable is scaled on the range $[10, 40]$ via the transformation $ax + b$, with $a = 30/5.3$ and $b = 40 - 4a$.

scores and class sizes, A and B, where both samples are drawn from the same true model 2.6.12. Suppose further that the observed class size range in sample A is $[10, 40]$, while in sample B it is $[11, 40]$. This difference might arise for example due to a difference in the shares of public expenditure for education in two different countries.[15] Pictures 2.21a- show standard linear OLS estimates[16]. While the estimated slope in sample A is positive and

2.21b

Figure 2.20: Misspecified model on 2 different ranges



(a) Range A

(b) Range B

significant, the estimate in sample B is negative and insignificant. The main implication of this simulated example is that even if the true relationship of class size and test scores is *exactly* the same, model misspecification together with differences in the observed distributions of class size can induce differences in sign and significance of the estimates.

2.7 Conclusion

I analyzed the effect of class size on test scores. The results indicate that the effect of class size is non-monotone. I demonstrated how a non-monotone causal effect together with differences in the ranges of observed class sizes can potentially explain the controversy that exists within the literature. I derived implications for economic theory develop a simple model for an educational production function that can generate non-monotone a causal effect. Further, I developed a framework in which a broad class of penalized minimum distance estimators is revealed as projections of the constrained set on the constrained set. I utilized this result to find sufficient conditions for the consistency of constrained estimators for a broad class of constraints in two nonparametric IV estimation approaches: a kernel and a sieve approach. The sufficient conditions were shown to be related to the smoothness of the model regression function and to the rate of ill-posedness of the inverse problem. I also

[15] Class sizes differ between countries, Averett and McLennan (2004).

[16] For simplicity, there is no randomness and endogeneity in this example

showed that monotonically constrained and unconstrained Tikhonov estimators coincide when the model regression is an inner point of the set of monotone functions.

2.8 Appendix

2.8.1 Data generation process (DGP) of the simulation (estimation part)

The (DGP) has the following steps: 1) generate an independent stochastic variation: $U_1, \ldots, U_n \overset{iid}{\sim} N(0,1)$, 2) generate the regression error: $\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} N(0,1)$, 3) generate the exogenous instrument W: first simulate the unscaled variable W_u with $W_{u,1}, \ldots, W_{u,n} \overset{iid}{\sim} N(0,1)$, then scale it to the interval $[-1,1]$: $W_i = \frac{W_{u,i} - min(W_{u,1},\ldots,W_{u,n})}{max(W_{u,1},\ldots,W_{u,n}) - min(W_{u,1},\ldots,W_{u,n})}$, 4) generate an endogenous regressor: $X_{k,i} = \beta_1\epsilon_i + \beta_2 W_{u,i} + \beta_3 U_i$, 5) transform its distribution so that it is uniformly distributed (insures that we have enough observations at the boundary): $X_{u,i} = F(X_{k,i})$, where F is the c.d.f of $N(0, \beta_1^2+\beta_2^2+\beta_3^2)$, 6) finally, rescale the uniform regressor to $[-1,1]$ according to the rule $X_i = \frac{X_{u,i} - min(X_{u,1},\ldots,X_{u,n})}{max(X_{u,1},\ldots,X_{u,n}) - min(X_{u,1},\ldots,X_{u,n})}$. This is necessary because I choose the Legendre Polynomial basis in L_2.

2.8.2 Proof of propositions

Proof of proposition 2.4.1. Define $\bar{f} = \sup_{x\in[0,1]}\{f(x)\}$ and $\underline{f} = \inf_{x\in[0,1]}\{f(x)\}$. Using the Sobolev's Inequality (see e.g Agmon (1965), p.32), there exist positive constants γ and r and a function $\widetilde{m}_n \in H^2_X$ with $\widetilde{m}_n = \widehat{m}_n$ almost everywhere, such that $\widetilde{m}_n \in C^1[0,1]$ and

$$| \widetilde{m}_n'(x) - m_0'(x) | \leq \gamma(\|\widetilde{m}_n - m_0\|_{H^2_X} + r\|\widetilde{m}_n - m_0\|_{L^2_X}). \tag{2.8.1}$$

Because of $\|\widehat{m}_n - m_0\|_{H^2_X} = o_p(1)$ we obtain

$$\|\widetilde{m}_n - m_0\|_{H^2_X} = o_p(1) \qquad \text{and trivially} \tag{2.8.2}$$

$$\|\widetilde{m}_n - m_0\|_{L^2_X} = o_p(1), \tag{2.8.3}$$

and hence

$$\|\widetilde{m}_n'(x) - m_0'(x)\|_\infty < c. \tag{2.8.4}$$

with probability tending to one. □

Proof of proposition 2.4.4.1. Following the lines of proof of Mammen and Thomas-Agnan (1999), let $m \in H$ and for some $\beta \in \mathbb{R}$ let $g = \widehat{m} + \beta(\widehat{m} - m)$. Since $\widehat{m}$ is a minimizer of

$\|Fm - h\|_G^2 + \alpha\|m\|_H^2$ over H, it must hold

$$\|F\widehat{m} - h\|_G^2 + \alpha\|\widehat{m}\|_H^2 \leq \|Fg - h\|_G^2 + \alpha\|g\|_H^2, \tag{2.8.5}$$

where after substituting for g the right hand side can be written as

$$\begin{aligned}
&\|F(\widehat{m} + \beta(\widehat{m} - m)) - h\|_G^2 + \alpha\|(\widehat{m} + \beta(\widehat{m} - m))\|_H^2 = \|F\widehat{m} - h\|_G^2 + \alpha\|\widehat{m}\|_H^2 + \\
&+2\beta\Big(< F\widehat{m} - h, F(\widehat{m} - m) >_G + \alpha < \widehat{m} - m, \widehat{m} >_H \Big) + \\
&+\beta^2\Big(\|F(\widehat{m} - m)\|_G^2 + \alpha\|(\widehat{m} - m)\|_H^2\Big),
\end{aligned}$$

so that inequality 2.8.5 is equivalent to

$$\begin{aligned}
&0 \leq 2\beta\Big(< F\widehat{m} - h, F(\widehat{m} - m) >_G + \alpha < \widehat{m} - m, \widehat{m} >_H \Big) + \\
&+\beta^2\Big(\|F(\widehat{m} - m)\|_G^2 + \alpha\|(\widehat{m} - m)\|_H^2\Big) = \beta A_1 + \beta^2 A_2
\end{aligned}$$

Dividing both sides by β^2 and letting $\beta \to 0$ from the left and from the right yields $A_1 = 0$. Now define

$$R := \|Fm - h\|_G^2 + \alpha\|m\|_H^2 - \|F(\widehat{m} - m)\|_G^2 - \alpha\|(\widehat{m} - m)\|_H^2.$$

After throwing out equal terms with opposite signs we obtain

$$\begin{aligned}
&R = 2 < Fm, F\widehat{m} >_G - 2 < Fm, Fh >_G + < h, h >_G - \\
&- < F\widehat{m}, F\widehat{m} >_G + 2\alpha < \widehat{m}, m >_H - \alpha < \widehat{m}, \widehat{m} >_H = \\
&= < F\widehat{m}, F\widehat{m} >_G + < h, F\widehat{m} >_G + < F\widehat{m}, Fm >_G - < Fm, h >_G - \\
&- \alpha < \widehat{m}, \widehat{m} >_H + \alpha < m, \widehat{m} >_H - \\
&- < h, F\widehat{m} >_G + < F\widehat{m}, Fm >_G - < Fm, h >_G + \alpha < m, \widehat{m} >_H = \\
&- A_1 + < F\widehat{m} - h, Fm >_G - < h, F\widehat{m} >_G + \alpha < m, \widehat{m} >_H = \\
&= -A_1 - A_1 + \|F\widehat{m} - h\|_G^2 + \alpha\|\widehat{m}\|_H^2 = \|F\widehat{m} - h\|_G^2 + \alpha\|\widehat{m}\|_H^2
\end{aligned}$$

and the last expression does not depend on m. □

Proof of corollary 2.4.4.2. It holds

$$\|\widehat{m}_C - m_0\|_{\mathscr{V}} \leq \|\widehat{m}_C - \widehat{m}\|_{\mathscr{V}} + \|\widehat{m} - m_0\|_{\mathscr{V}} \leq \|\widehat{m} - m_0\|_{\mathscr{V}} + \|\widehat{m} - m_0\|_{\mathscr{V}}. \tag{2.8.6}$$

If additionaly C is convex and closed, then it holds

$$\|\widehat{m}_C - m_0\|_{\mathscr{V}} = \|\Pi_C \widehat{m} - \Pi_C m_0\|_{\mathscr{V}} \leq \|\widehat{m} - m_0\|_{\mathscr{V}}, \tag{2.8.7}$$

where Π_C is the projection on the set C. □

For the proof of proposition 2.4.5.1 I need to prove first the following Lemma:

Lemma 2.8.2.1. *(A modification of the Poincaré's inequality) Let ϕ be bounded and nonnegative on $[0,1]$, with $\int_0^1 \phi(u)du = 1$. Then it holds for m_1, m_2 in H and a positive constant c*

$$\|m_1 - m_2\|^2_{L^2_{([0,1])}} \leq c\Big(\|m_1' - m_2'\|^2_{L^2_{([0,1])}} + \int_0^1 (m_1(x) - m_2(x))^2 \phi(x)dx\Big). \tag{2.8.8}$$

Proof of Lemma 2.8.2.1. It suffices to show the inequality for functions m_j with $m_j(x) = m_j(u) + \int_u^x m_j'(y)dy$. As $\int_0^1 \phi(u)du = 1$, we have

$$m_j(x) = \int_0^1 m_j(u)\phi(u)du + \int_0^1 \Big(\phi(u)\int_u^x m_j'(y)dy\Big)du$$

It follows that

$$\begin{aligned}
&|\int_0^1 (m_1(x) - m_2(x))dx| \leq \int_0^1 |(m_1(x) - m_2(x))|dx \leq \\
&|\int_0^1 (m_1(u) - m_2(u))\phi(u)du| + \int_0^1 |\int_0^1 \phi(u)\int_u^x (m_1'(y) - m_2'(y))dydu|dx \\
&\leq \int_0^1 (m_1(u) - m_2(u))^2\phi(u)du + \|m_1' - m_2'\|^2_{L^2_{([0,1])}},
\end{aligned}$$

where the last inequality follows from a multiple application of the Cauchy-Schwarz inequality. □

Proof of Proposition 2.4.5.1. Let $(\overrightarrow{\widetilde{m_p}})$ be a Cauchy sequence in $(V, \|\|_V)$. The Cauchy property of $(\overrightarrow{\widetilde{m_p}})$ implies that for each $i = 1, \ldots, n$ the sequence $(m_{pi}(X_i))$ is a Cauchy-Sequence in $\mathbb{R}$ and the sequence (m_{pi}') is a Cauchy sequence in $L^2_{([0,1])}$. Denote with $m_{0i}(X_i)$ the limit of $(m_{pi}(X_i))$ in $\mathbb{R}$ and with $m_{0,i}'$ the limit of (m_{pi}') in $L^2_{([0,1])}$. The proof follows using Lemma 2.8.2.1 with $\phi(u) = \frac{1}{n}K\Big(\frac{X_i - u}{h}\Big)$. □

Proof of proposition 2.4.2. Define $< .,. >_Q$: $H \times H \to \mathbb{R}_+$, $(l, g) \to < l, g >_Q := n^{-1}\sum_{i=1}^n l(W_i)g(W_i)$. Then this is a positive semidefinite bilinear form and $\widehat{m}$ is a solution to $\min_{m\in G} < \widehat{T}m - \widehat{h}, \widehat{T}m - \widehat{h} >_Q + \alpha_n P(m)$. Observe that the proof of proposition 2.4.4.1 only uses the properties of

a positive semidefinite bilinear form (i.e. positive definiteness is not needed). Therefore, following identical steps as in its proof, we obtain

$$\widehat{m}_C \in \underset{m \in C}{\operatorname{argmin}} < \widehat{T}(\widehat{m} - m), \widehat{T}(\widehat{m} - m) >_Q + \alpha_n P(\widehat{m} - m), \tag{2.8.9}$$

and using the Cauchy-Schwarz inequality we obtain

$$\|\widehat{T}(\widehat{m}_C - m_0)\|_Q^2 + \alpha_n P(\widehat{m}_C - m_0) \leq 2(\|\widehat{T}(\widehat{m} - m_0)\|_Q^2 + \alpha_n P(\widehat{m} - m_0)), \tag{2.8.10}$$

with $\|g\|_Q^2 := < g, g >_Q$.

Next, define $\bar{f} := \sup_{x \in [0,1]} \{| f_X(x) |\}$ and $\underline{f} := \inf_{x \in [0,1]} \{| f_X(x) |\}$. With assumption 3 ii), $0 < \underline{f} \leq \bar{f} < \infty$. With the definition of the penalty P, we have $\underline{f} P(m) \leq \|m\|_G^2 \leq \bar{f} P(m)$. Under assumptions 1-5, $\|\widehat{m} - m_0\|_G = o_p(1)$ due to and with $< \widehat{T}(\widehat{m} - m_0), \widehat{T}(\widehat{m} - m_0) >_Q = o_p(\alpha_n)$ we obtain

$$< \widehat{T}(\widehat{m} - m_0), \widehat{T}(\widehat{m} - m_0) >_Q + \alpha_n P(\widehat{m} - m_0) = o_p(\alpha_n). \tag{2.8.11}$$

Finally, using 2.8.10 and 2.8.11, we have $P(\widehat{m}_C - m_0) = o_p(1)$ and consequently $\|\widehat{m}_C - m_0\|_G = o_p(1)$. □

Proof of proposition 2.4.3. In a first step, we write

$$\begin{aligned}
&n^{-1} \sum_{i=1}^{n} (\widehat{T}(\widehat{m} - m_0)(W_i))^2 = n^{-1} \sum_{i=1}^{n} (\widehat{T}\widehat{m}(W_i) - T\widehat{m}(W_i) + T\widehat{m}(W_i) - Tm_0(W_i) \\
&+ Tm_0(W_i) - \widehat{T}m_0(W_i))^2 \leq \\
&3n^{-1} \Big(\sum_{i=1}^{n} (\widehat{T}\widehat{m}(W_i) - T\widehat{m}(W_i))^2 + \sum_{i=1}^{n} (T\widehat{m}(W_i) - Tm_0(W_i))^2 + \\
&\sum_{i=1}^{n} (Tm_0(W_i) - \widehat{T}m_0(W_i))^2 \Big) = A_1 + A_2 + A_3.
\end{aligned}$$

To bound A_2, because $T\widehat{m} - Tm$ suffices the conditions of the uniform law of large numbers, we have

$$n^{-1} \sum_{i=1}^{n} (T\widehat{m}(W_i) - Tm_0(W_i))^2 \to \|T\widehat{m} - Tm_0\|_W^2 \quad \text{almost surely.} \tag{2.8.12}$$

Further, we have $\|T\widehat{m} - Tm_0\|_W^2 \leq \|T\| \|\widehat{m} - m_0\|_X^2$, where $\|.\|$ is the operator norm. With a choice $\alpha_n \asymp \delta_{m,n}^{*2} \sqrt{\phi(v_k^{-2})}$, which in the case $\phi(x) = x^s$ corresponds to $\alpha_n \asymp \delta_{m,n}^{*1+\frac{s}{\alpha+s}}$, and under the assumptions 1-8, corollary 5.2 in Chen and Pouzo (2012) holds and

$$\|\widehat{m} - m_0\|_X = O_p(\delta_{m,n}^{*\frac{\alpha}{\alpha+s}}). \tag{2.8.13}$$

Consequently,

$$\frac{\|\widehat{m}-m_0\|_X^2}{\alpha_n} = \frac{\|\widehat{m}-m_0\|_X^2}{\delta_{m,n}^{*\frac{2\alpha}{\alpha+s}}}\frac{\delta_{m,n}^{*\frac{2\alpha}{\alpha+s}}}{\alpha_n} = O_p(1)\delta_{m,n}^{*\frac{\alpha-2s}{\alpha+s}}. \tag{2.8.14}$$

Therefore, if $\alpha > 2s$ then $n^{-1}\sum_{i=1}^n (T\widehat{m}(W_i) - Tm_0(W_i))^2 = o_p(\alpha_n)$.

To bound A_3, observe that

$$n^{-1}\sum_{i=1}^n (Tm_0(W_i) - \widehat{T}m_0(W_i))^2 = n^{-1}\sum_{i=1}^n (h(W_i) - \widehat{h}(W_i) + \widehat{h}(W_i) - \widehat{T}m_0(W_i))^2 \leq$$
$$2n^{-1}\sum_{i=1}^n (h(W_i) - \widehat{h}(W_i))^2 + 2n^{-1}\sum_{i=1}^n (\widehat{h}(W_i) - \widehat{T}m_0(W_i))^2 = A_{3,1} + A_{3,2}.$$

Because $m_0 \in \mathscr{H}_{os}$, then under assumptions 1-5, lemma C.2 in Chen and Pouzo (2012) holds. Therefore $A_{3,2} \leq O_p(\delta_n^2)$ where $\delta_n^2 = \max\{J_n/n, b_{J_n}^2\}$. Because the estimators of T and h are least squares series estimators, it holds $\delta_n^2 = J_n/n \asymp b_{J_n}^2$, and with $\alpha \asymp \delta_{m,n}^{*1+\frac{s}{\alpha+s}}$ we obtain $A_{3,2} = o_p(\alpha_n)$. To bound $A_{3,1}$, note that $\widehat{h}$ suffices the conditions of the uniform law of large numbers. Then, using the same line of reasoning as for A_2, we obtain that $A_{3,1} = O(b_{J_n}^2)$ and is hence equal to $o_p(\alpha_n)$.

To bound A_1, observe that

$$n^{-1}\sum_{i=1}^n (\widehat{T}\widehat{m}(W_i) - T\widehat{m}(W_i))^2 \leq n^{-1}\sum_{i=1}^n (\widehat{T}-T)(\widehat{m}(W_i) - m_0(W_i))^2 + \frac{2}{3}A_3.$$

Following the line of reasoning of the previous point, we achieve $A_1 = o_p(\alpha_n)$. □

Proof of 2.4.4. From the projection framework, corollary 2.4.4.2, we have

$$\int |\,\widehat{T}(\widehat{m}_C(w) - m_0)(w)\,|^2\,\lambda(dw) + \alpha_n\|\widehat{m}_C - m_0\|_{H^1}^2 \leq \tag{2.8.15}$$
$$\int |\,\widehat{T}(\widehat{m}(w) - m_0)(w)\,|^2\,\lambda(dw) + \alpha_n\|\widehat{m} - m_0\|_{H^1}^2.$$

Under assumptions K1-K6, $\alpha_n\|\widehat{m}-m_0\|_{H^1}^2 = o(\alpha_n)$. With 2.4.21, the right-hand-side of inequality 2.8.15 is equal to $o_p(\alpha_n)$, and therefore $\alpha_n\|\widehat{m} - m_0\|_{H^1}^2 = o(\alpha_n)$. □

Proof of proposition 2.4.5. It holds

$$\|\widehat{T}(\widehat{m} - m_0)\|_H^2 = \|\widehat{T}\widehat{m} - T\widehat{m} + T\widehat{m} - Tm_0 + Tm_0 - \widehat{T}m_0\|_H^2 \tag{2.8.16}$$
$$\leq 3\Big(\|\widehat{T}\widehat{m} - T\widehat{m}\|_H^2 + \|T\widehat{m} - Tm_0\|_H^2 + \|Tm_0 - \widehat{T}m_0\|_H^2\Big) \tag{2.8.17}$$
$$\leq 3\Big(\|\widehat{T}-T\|^2\|\widehat{m}\|_G^2 + \|T\|^2\|\widehat{m} - m_0\|_G^2 + \|\widehat{T}-T\|^2\|m_0\|_G^2\Big) \tag{2.8.18}$$
$$= 3\Big(\|\widehat{T}-T\|^2(\|\widehat{m}\|_G^2 + \|m_0\|_G^2) + \|T\|^2\|\widehat{m} - m_0\|_G^2\Big). \tag{2.8.19}$$

Since $\|\widehat{m} - m_0\| = o_p(1)$, $\|\widehat{m}\|_G^2 + \|m_0\|_G^2$ is bounded in probability. Under the assumptions of this proposition, $\|\widehat{T} - T\|^2 = O_p(\frac{1}{n\sigma^2} + \sigma^{2\rho})$ and

$$\|\widehat{m} - m_0\|^2 = O_p\Big[\frac{1}{\alpha_n^2}\Big(\frac{1}{n} + \sigma^{2\rho}\Big) + \Big(\frac{1}{n\sigma^2} + \sigma^{2\rho}\Big)\alpha_n^{\min(\beta-1,0)} + \alpha_n^{\min(\beta,2)}\Big], \tag{2.8.20}$$

see theorem 4.1 in Darolles, Fan, Florens, and Renault (2011) and the supplement materials of Darolles, Fan, Florens, and Renault (2011). Following their considerations, a choice $\alpha_n \propto n^{-1/(\min(\beta,2)+2)}$ leads to the equivalence

$$\frac{1}{n\alpha_n^2} \sim \alpha_n^{\min(\beta,2)} \sim n^{-(\min(\beta,2)/(\min(\beta,2)+2)}, \tag{2.8.21}$$

and to $\Big(\frac{1}{n\sigma^2} + \sigma^{2\rho}\Big)\alpha_n^{\min(\beta-1,0)} = O\Big(\frac{\alpha_n^{\min(\beta-1,0)}}{n\sigma^2}\Big)$, and in particular, under a suitable choice for σ, this implies $\frac{1}{n\sigma^2} = O\Big(\frac{\alpha_n^{\min(\beta,2)}}{\alpha_n^{\min(\beta-1,0)}}\Big)$.

Next, suppose $\beta > 1$. Then

$$\frac{\|\widehat{T} - T\|^2}{\alpha_n} = \frac{\|\widehat{T} - T\|^2}{\frac{1}{n\sigma^2} + \sigma^{2\rho}} \frac{\frac{1}{n\sigma^2} + \sigma^{2\rho}}{\alpha_n} = O_p(1)o(1) = o_p(1). \tag{2.8.22}$$

Further, observe that with this choice of α_n, the rate of convergence $O_p\Big[\frac{1}{\alpha_n^2}\Big(\frac{1}{n} + \sigma^{2\rho}\Big) + \Big(\frac{1}{n\sigma^2} + \sigma^{2\rho}\Big)\alpha_n^{\min(\beta-1,0)} + \alpha_n^{\min(\beta,2)}\Big]$ is determined by $\alpha_n^{\min(\beta,2)}$ (due to the considerations described above). With $\beta > 1$, it follows that $\|\widehat{m} - m_0\|^2 = o_p(\alpha_n)$. With $\|\widehat{T} - T\|^2 = o_p(\alpha_n)$ and $\|\widehat{m} - m_0\|^2 = o_p(\alpha_n)$ in plugged in 2.8.19, we obtain the wished result. □

2.8.3 The Maimonides' rule: the study of Angrist and Lavy (1999)

As an illustration, I provide evidence for nonmonotone causal effect in a second data set. The paper of Angrist and Lavy (1999) studies the effects of class size on test scores using an administrative class size cap rule as an instrument for the endogenous class size. The paper had a considerable influence on the literature: after it was published in 1999, a number of other studies used its identification strategy, see for example Hoxby (2000), Dobbelsteen, Levin, and Oosterbeek (2002), Gary-Bobo and Mahjoub (2006) and Urquiola (2006). The dataset in Angrist and Lavy (1999) comes from Jewish public secular primary schools. In 1990-1991 fourth graders and fifth graders and in 1991-1993 third graders were evaluated in a verbal and a math tests. The tests were organized by a testing center under the Israel Ministry of Education.[17] The data set contains, amongst others, school identifier, total enrollment for each school, class size for each class, average test score in each class, as well as the school-level socio-economic variable percentage disadvantaged (PD) students. The unit of observation is

[17]The test was conducted in all primary schools. The sample of the study is restricted to secular Jewish schools.

the class.[18] The benchmark model of Angrist and Lavy (1999) is defined as

$$Y_{sc} = X_s'\beta + n_{sc}\alpha + \eta_s + \epsilon_{sc}, \tag{2.8.23}$$

where Y_{sc} is the average test score in class c for school s in a particular subject, X_s is a vector of school characteristics, n_{sc} is the class size, η_s is a random school component and ϵ_{sc} is a class level error term. The identification strategy is based on the following administrative rule, referred to as the Maimonides' rule: when in a cohort of p classes the class size of 40 students is exceeded, one additional class is added and the cohort is split into $p + 1$ classes of equal size. Formally, the class size f_{sc} assigned to class c in school s for a specific grade is determined by the formula

$$f_{sc} = e_s/[1 + (e_s - 1)/40], \tag{2.8.24}$$

where e_s is the enrollment for that school and grade at the beginning of the school year and for any $r > 0$ $[r]$ denotes the largest integer not exceeding r. The rule is illustrated on figure 2.21. The x-axes is for enrollment size and the y-axis for the class size. The dashed line depicts the

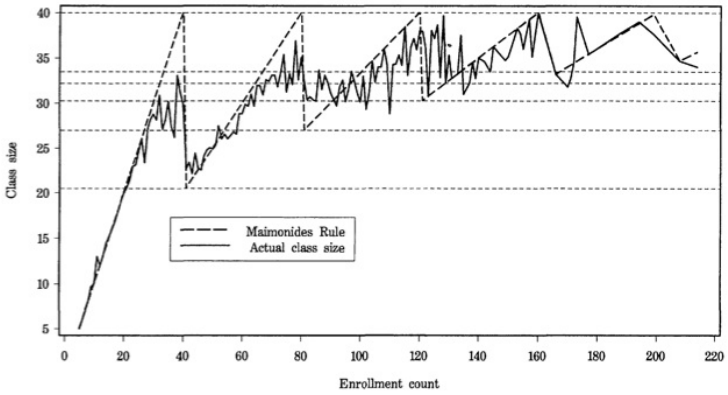


Figure 2.21: Maimonides' Rule. Source: Angrist and Lavy (1999).

assigned class size as derived by the Maimonides' rule and the continuous line for effectively implemented class size. The due-class size function has jumps at 41, 81, 121, and so forth. Angrist and Lavy (1999) exploit the discontinuities of the assignment rule in a fuzzy regression discontinuity design (RDD) to identify the causal effect. The method goes back to Trochim (1984) and the connection to 2SLS was unveiled by van der Klaauw (2002). In the standard approach with a binary treatment, there is a jump in the probability of the binary treatment at the cut-off point of the forcing variable. If all other factors vary smoothly with the forcing variable, a jump in the average outcome at the cut-off point is interpreted as a treatment effect. Angrist and Lavy (1999) modify the standard approach to a case with a multiple-valued treatment (here the class size) and multiple cut-off points of the forcing variable (here the enrollment). As a result, they exploit a change in the average class size and estimate an average of the treatment effects at three different discontinuities. Their estimations rely

[18]The data for third graders is on individual level but is not available and therefore I restrict the analysis in my paper on fourth and fifth graders. The dataset can be downloaded from the homepage of Joshua Angrist, http://economics.mit.edu/faculty/angrist/data1.

on a parametric specification of the trend function that consists of smooth covariates, most importantly of enrollment. A potential problem associated with parametric specifications in RDD context is that nonlinearities in the regression function might be mistakenly estimated as a jump (and hence, as a treatment effect). Angrist and Lavy (1999) address this problem with a robustness check, in which they only use the observations close to the discontinuities, that is, in the intervals $[36, 45], [76, 85], [116, 125]$ and, alternatively, $[38, 43], [78, 83], [118, 123]$. The authors refer to these observations as to the discontinuity sample. Although the choice of the interval length is arbitrary, narrowing the sample to observations close to the discontinuities reduces the potential danger for misspecification, provided the regression function is smooth.

The estimates of α are negative in all cases but insignificant for the 3rd graders, most of the cases for the 4th graders and some of cases for the 5th graders.

A potential source of model specification error is that the treatment effect is assumed to be linear in the class size. This amounts to estimating a (weighted) average of the treatment effects at the different discontinuity points. The main advantage of such an aggregation when the estimation is done only with observation from the discontinuity sample, is that the observations from all three intervals can be used. Around each cut-off point there are only few points. For example, for fourth graders there are about 60 observations in the -3/+3-interval around the cut-off point 120, that is, only 30 observations above and below the cut-off point.[19] Estimating different treatment effects would be associated with high imprecision. The linearity assumption makes it possible to pool the observations together. The price to pay for the higher precision is a potential model misspecification. The following example illustrates a model specification in the context of Cho, Glewwe, and Whitler (2012). Consider the estimated cubic polynomial for 5th grade math in the data set of Cho, Glewwe, and Whitler (2012), depicted on figure 2.18a. Suppose that this function summarizes the true causal effect of class size on test scores, and that it is estimated using the Maimonides' rule under the assumption that the regression function is linear in the treatment. The estimated treatment effect will be a weighted average of the negative effects at the points 40, 80 and 120. Thus, the single estimate will be negative and the information on the magnitude of the three effects is lost. The non-monotonicity of the treatment effect cannot be detected.

An average class size approach

The main restriction of an empirical strategy based on a maximum class-size rule is that it provides identification of the treatment effect only at the discontinuities. An alternative approach is to use the average class size as an instrument. Mainly because of its availability, this approach has been often explored in the literature, see for example the studies of Akerhielm (1995) and Bressoux, Kramarz, and Prost (2009). Another important advantage of this instrument is that it can potentially explain variation in the endogenous regressor on almost the whole range of observed class sizes.

If the only source of endogeneity is the within-school selection, then the average class size is potentially a valid instrument. As between-schools selection generally cannot be excluded, using the whole sample might bias the estimates. The main threat for the validity of the exclusion restriction is the selective behavior of parents (selection of teachers apples similarly). Parents who invest more in the human capital of their children might seek to get

[19] One of the reasons for this phenomenon is that schools split classes at enrollments that are usually smaller than 40, 80 and so forth.

them into schools with on average smaller class sizes. A crucial restriction in the context of the school system in Israel is that parents are not completely free to move their children to different public schools. Once the parents have chosen their residence, they can only give their children to the school in whose catchment area they live.[20] Hence, the main endogeneity concern is that prior to begin of the schooling, parents choose their residence so as to maximize the educational quality for their children given a budget constraint. Moreover, although investments in the human capital of the children are not observed in the data, it is considered to be strongly predicted by the socio-economic status of the parents. One crucial implication is that the socio-economic composition in a school is likely to be highly correlated with the *initial* sorting of the parents. Therefore, comparing schools with a similar socio-economic composition potentially solves the problem of between-schools selection. Although there is no complete information on this composition, the variable PD provides a proxy for it. For this reason I restrict the analysis that follows on subsamples of schools with similar or equal values of the PD variable.[21]

Although the exclusion restriction cannot be tested, it can be addressed indirectly in a framework first developed by Horowitz (2012). Suppose W is a valid instrument with $E[\epsilon|W] = 0$ and that the researcher is interested in testing the hypothesis that there exists a smooth function m_0 that satisfies $Y = m_0(X) + \epsilon$. That is, the null hypothesis is defined as

$$H_s: \quad \text{There exists a smooth and additively separable in } X \text{ and } \epsilon \text{ function } m_0 \text{ that satisfies } Y = m_0(X) + \epsilon.$$

The test statistic has a normal distribution under the null. High values of the test statistic indicate a model specification error. This error can also be due to a violation of the exclusion restriction. I implement a modification of this test developed by Breunig (2012) for different values of the PD (3, 5) variables. I use the assigned class size f_{sc} as a potential instrument instead of the actual class size. The main reason is that the assigined class size depends solely on the total enrollment in a school, whereas the average class size can be manipulated via splitting classes or pulling classes together. The resulting p-values vary in the range of 0.3 to 0.8. Thus, there is no evidence for model misspecification eror in the context of the null 2.8.25. At least indirectly, this outcome supports the plausibility of the exclusion restriction.

Testing for monotonicity

As in the previous application, my main focus is to explore the functional form of the causal effect of class size on test scores. In a first step, I test the hypothesis that the causal effect is monotone. Then the null hypothesis is

$$H_0 : m' \leq 0.^{22} \tag{2.8.25}$$

The cutoff parameter and the regularization constant of the constrained estimator are chosen as in the previous empirical application. The results are shown in table 2.20. The first col-

[20] As noted by Angrist and Lavy (1999), private schools are not common in Israel outside the Jewish ultra-Orthodox sector.

[21] Although Angrist and Lavy (1999) acknowledge the importance of this variable, they use it in a different context.

[22] I discuss only the null for decreasing function, the other case goes similarly.

Table 2.20: p-Values, Test for Monotonicity

m	n	verbal	math	n	verbal	math
PD ≤ 3						
4	410	0.0001	0	401	0.006	0.058
5	410	0.0002	0	401	0	0.11
6	410	0.001	0	401	0	0.0003
7	410	0.0002	0	401	0	0
PD ≤ 5						
4	682	0.002	0.001	661	0.0002	0.564
5	682	0.006	0.0099	661	0	0.18
6	682	0.001	0	661	0	0
7	682	0	0	661	0	0
8	682	0	0	661	0	0

umn contains information about the cutoff parmeter and the second one about the sample size. The first part of the table presents the results of the test with the sample with PD ≤ 3 unconditionally. p-values that are smaller than 0.00001 were reported as zero. For all values of the cutoff parameter in the range 4-7 the test rejects the monotonicity. Conditioning on PD ≤ 5 yields similar results. The p-values are somewhat higher but still below 1%. The only exceptions are the p-values for the 5th grade math regression in the subgoup PD ≤ 5 for cutoff values 4 and 5. Altogether, if the exlcusion restriction is valid, then there is a strong evidence for a nonmonotone causal effect of class size on test scores.

Parametric forms of the regression function

Following the line of arguments in subsection 2.6, I explore the form of the regression function by testing for parametric specification. As in the previous application, I restrict my search to polynomials of second and third degree and I follow the same structure as in the case of testing for monotonicity. The results are shown in table 2.21. Column one contains the value of the cutoff parameter m of the test statistic, column 2 the sample size. All columns with a header value $d = 2$ contain the p-values for a quadratic specification and those with $d = 3$ for a cubic one. For 5th grade math and for higher values of the cutoff parameter (8,9,10) the p-values are between 0.03 and 0.08. For PD ≤ 3 the p-values are similar for all specifications and higher than 10 % with the exception of the 5th grade verbal case, wher the test rejects the quadratic specification for higher values of m_n. For PD ≤ 5 both quadratic and cubic specifications yield high p-values for low values of m_n and low p-values for high values of m_n. An exception is again the case of 5th grade verbal where the cubic regression

yield high p-values and the quadratic one is rejected. When I test only with the restricted sample, the p-values are always higher than 10 %.[23] Based on this evidence both quadratic and cubic polynomials represent a plausible functional for the regression function with the cubic polynomial being in about 25 % of the cases more robust than the quadratic one.

[23] With the restricted sample I only condition on PD ≤ 10 as the subgroups PD ≤ 3 and PD ≤ 5 does not contain enough observations.

Table 2.21: p-Values, Test for Parametric Specification, dataset of Angrist and Lavy (1999)

		4th grade verbal		4th grade math			5th grade verbal		5th grade math	
m	n	d=2	d=3	d=2	d=3	n	d=2	d=3	d=2	d=3
PD ≤ 3										
4	410	0.16	0.16	0.19	0.16	401	0.4	0.16	0.16	0.16
5	410	0.17	0.17	0.20	0.26	401	0.03	0.27	0.31	0.24
6	410	0.13	0.13	0.73	0.95	401	0.0001	0.85	0.50	0.67
7	410	0.28	0.29	0.50	0.43	401	0.0003	0.83	0.25	0.37
PD ≤ 5										
4	682	0.17	0.16	0.26	0.16	661	0.33	0.16	0.19	0.15
5	682	0.13	0.11	0.20	0.21	661	0.1	0.22	0.03	0.35
6	682	0.40	0.31	0.69	0.34	661	0	0.78	0.0002	0.02
7	682	0.16	0.16	0	0	661	0	0.26	0	00
8	682	0.006	0.008	0	0	661	0	0.25	0	0

Chapter 3

Bibliography

Abbring, J. H., and J. J. Heckman (2007): "Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6 of *Handbook of Econometrics*, chap. 72. Elsevier.

Abbring, J. H., and G. J. van den Berg (2003): "The non-parametric identification of treatment effects in duration models," *Econometrica*, 71(5), 1491–1517.

——— (2005): "Social experiments and Instrumental Variables with Duration Outcomes," Tinbergen Institute Discussion Paper 05 - 047/3, Tinbergen Institute, The Netherlands.

Agmon, S. (1965): *Lectures on elliptic boundary value problems*. Van Nostrand Mathematical Studies.

Akerhielm, K. (1995): "Does class size matter?," *Economics of Education Review*, 14(3), 229–241.

Andersen, P., Ø. Borgan, R. Gill, and N. Keiding (1997): *Statistical Models Based on Counting Processes*, Springer Series in Statistics. Springer New York.

Andrews, D. W. K. (1994): "Empirical process method in econometrics," Handbook of Econometrics. Cambridge, MA: Ballinger.

Angrist, J. D., and V. Lavy (1999): "Using Mammonites' rule to estimate the effect of class Size on Scholastic Achievement," *The Quarterly Journal of Economics*, 114(2), 533–575.

Averett, S. L., and M. C. McLennan (2004): "Exploring the effect of class size on student achievement: what have we learned over the past two decades?," in *International handbook on the economics of education*, ed. by G. Johnes, and J. Johnes, chap. 9, pp. 329–367. Edward Elgar Publishing Limited.

Bergemann, A., M. Caliendo, G. J. van den Berg, and K. Zimmermann (2011): "The threat effect of participation in active labor market programs on job search behavior of migrants in Germany," *International Journal of Manpower*, 7, 777–795.

Bijwaard, G. (2008): "Instrumental variable estimation for duration data," Tinbergen Institute Discussion Paper 08-032/4, Tinbergen Institute, The Netherlands.

Bijwaard, G., and G. Ridder (2005): "Correcting for selective compliance in a re-employment bonus experiment," *Journal of Econometrics*, 125, 77–111.

Black, D. A., J. A. Smith, M. C. Berger, and B. J. Noel (2003): "Is the threat of reemployment services more effective than the services themselves? Evidence from random assignment in the UI system," *The American Economic Review*, 93(4), 1313–1327.

Blasco, S. (2009): "Do people forgo extra money to avoid job search assistance," Discussion paper, CREST, Paris, France.

Blatchford, P., P. Bassett, and H. Goldstein (2003): "Are class size differences related to pupils' educational progress and classroom processes? Findings from the Institute of Education class size sudy of children aged 5-7 years," *British Educational Research Journal*, 29(5), pp. 709–730.

Bloom, H. S. (1984): "Estimating the effect of job-training programs, using longitudinal data: Ashenfelter's findings reconsidered," *The journal of human resources*, 19, 544–556.

Bloom, H. S., L. L. Orr, S. H. Bell, G. Cave, F. Doolittle, W. Lin, and J. M. Bos (1997): "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study," *Journal of Human Resources*, 32(3), 549–576.

Blundell, R., X. Chen, and D. Kristensen (2007): "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," *Econometrica*, 75(6), 1613–1669.

Blundell, R., M. C. Dias, C. Meghir, and J. V. Reenen (2004): "Evaluating the Employment Impact of a Mandatory Job Search Program," *Journal of the European Economic Association*, 2(4), 569–606.

Bonesrøinning, H. (2003): "Class Size Effects on Student Achievement in Norway: Patterns and Explanations," *Southern Economic Journal*, 69(4), pp. 952–965.

Bonnal, L., D. Fougère, and A. Sérandon (1997): "Evaluating the impact of French employment policies on individual labour market histories," *The Review of Economic Studies*, 64(4), 683–713.

Boozer, M., and C. Rouse (2001): "Intraschool Variation in Class Size: Patterns and Implications," *Journal of Urban Economics*, 50(1), 163 – 189.

Bowman, A. W., M. C. Jones, and I. van der Gijbels (1998): "Testing monotonicity of regression," *Journal of Computational and Graphical Statistics*, 7, 489–500.

Bressoux, P., F. Kramarz, and C. Prost (2009): "Teachers' Training, Class Size and Students' Outcomes: Learning from Administrative Forecasting Mistakes," *Economic Journal*, 119(536), 540–561.

Breunig, C. (2012): "Goodness-of-fit tests based on series estimators in nonparametric instrumental regression," *Working paper*.

——— (2013): "Specification testing in nonparametric instrumental quantile regression," *Working paper*.

Brown, B. W., and D. H. Saks (1980): "Production technologies and resource allocations within classrooms and schools: theory and measurement," in *The Analysis of Educational Productivity*, ed. by R. Dreeben, and J. A. Thomas, Handbook of Econometrics, pp. 53–117. Cambridge, MA: Ballinger.

Case, A., and A. Deaton (1999): "School Inputs and Educational Outcomes in South Africa," *The Quarterly Journal of Economics*, 114(3), pp. 1047–1084.

Chen, X., O. Linton, and I. V. Keilegom (2003): "Estimation of Semiparametric models when the criterion function is not smooth," *Econometrica*, 71, 1591–1608.

Chen, X., and D. Pouzo (2012): "Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals," *Econometrica*, 80(1), 277–321.

Chesher, A. (2002): "Semiparametric identification in duration models," CeMMAP working paper CWP20/02, Centre for Microdata Methods and Practice, London, UK.

——— (2003): "Identification in Nonseparable Models," *Econometrica*, 71(5), pp. 1405–1441.

Cho, H., P. Glewwe, and M. Whitler (2012): "Do reductions in class size raise students' test scores? Evidence from population variation in Minnesota's elementary schools," *Economics of Education Review*, 31(3), 77 – 95.

Correa, H. (1993): "An economic analysis of class size and achievement in education," *Education Economics*, 1(3), 129 – 135.

Cox, D. (1962): *Renewal Theory*. Methuen, London, UK.

Crépon, B., M. Dejemeppe, and M. Gurgand (2005): "Counseling the unemployed: does it lower unemployment duration and recurrence?," IZA Discussion paper 1796, Institute for the Study of Labor (IZA), Bonn, Germany.

Crépon, B., M. Ferracci, and D. Fougère (2007): "Training the unemployed in France: how does it affect unemployment duration and recurrence?," IZA Discussion paper 3215, Institute for the Study of Labor (IZA), Bonn, Germany.

Crépon, B., M. Ferracci, G. Jolivet, and G. J. Van den Berg (2009): "Active labor market policy effects in a dynamic setting," *Journal of the European Economic Association*, 7(2-3), 595–605.

——— (2010): "Analyzing the anticipation of treatments using data on notification dates," IZA Discussion paper 5265, Institute for the Study of Labor (IZA), Bonn, Germany.

Currie, J. (2004): "The take-up of social benefits," Discussion paper, Conférence an l'honneur de Eugène Smolensky.

Darolles, S., Y. Fan, J. P. Florens, and E. Renault (2011): "Nonparametric Instrumental Regression," *Econometrica*, 79(5), 1541–1565.

Dobbelsteen, S., J. Levin, and H. Oosterbeek (2002): "The Causal Effect of Class Size on Scholastic Achievement: Distinguishing the Pure Class Size Effect from the Effect of Changes in Class Composition," *Oxford Bulletin of Economics and Statistics*, 64(1), 17–38.

Dormont, B., D. Fougère, and A. Prieto (2001): "L'effet de l'allocation unique dégressive sur la reprise d'emploi," *Économie et Statistique*, 343(1), 3–28.

Eberwein, C., J. C. Ham, and R. J. LaLonde (1997): "The impact of being offered and receiving classroom training on the employment histories of disadvantaged women: evidence from experimental data," *Review of Economic Studies*, 64(4), 655–682.

Engl, H., M. Hanke, and A. Neubauer (1996): *Regularization of Inverse Problems*, Mathematics and Its Applications. Springer.

Fitzenberger, B., O. Orlanski, A. Osikominu, and M. Paul (2013): "Déjà Vu? Short-term training in Germany 19801992 and 20002003," *Empirical Economics*, 44(1), 289–328.

Fredriksson, P., and P. Johansson (2008): "Dynamic Treatment Assignment," *Journal of Business & Economic Statistics*, 26, 435–445.

Freyssinet, J. (2002): "La réform de l'indemnisation du chômage en France," IRES Document de Travail 02.01, IRES.

Frölich, M. (2007): "Regression discontinuity design with covariates," IZA Discussion paper 3024, Institute for the Study of Labor (IZA), Bonn, Germany.

Gary-Bobo, R. J., and M. B. Mahjoub (2006): "Estimation of Class-Size Effects, Using 'Maimonides' Rule': The Case of French Junior High Schools," CEPR Discussion Papers 5754, C.E.P.R. Discussion Papers.

Gerfin, M., and M. Lechner (2002): "A microeconometric evaluation of the active labour market policy in Switzerland," *The Economic Journal*, 112(482), 854–893.

Gonzalez-Manteiga, W., and C. Cadarso-Suarez (2007): "Asymptotic properties of a generalized Kaplan-Meier estimator with some applications," *Journal of nonparametric statistics*, 4, 65–78.

Gorter, C., and G. R. J. Kalb (1996): "Estimating the effect of counseling and monitoring the unemployed using a job search model," *The Journal of Human Resources*, 31(3), 590–610.

Grasmair, M., O. Scherzer, and A. Vanhems (2013): "Nonparametric instrumental regression with non-convex constraints," *Inverse Problems*, 29(3), 035006.

Gritz, R. M. (1997): "The impact of training on the frequency and duration of employment," *Journal of Econometrics*, 57, 21–51.

Hahn, J., P. Todd, and W. van der Klaauw (2001): "Identification and estimation of treatment effects with a Regression-Discontinuity design," *Econometrica*, 69(1), 201–209.

Hall, P., and N. Heckman (2000): "Testing for monotonicity of a regression mean by calibrating for linear functions," *The Annals of Statistics*, 28, 20–39.

Ham, J. C., and R. J. LaLonde (1996): "The effect of sample selection and initial conditions in duration models: Evidence from experimental data on training," *Econometrica*, 64(1), 175–205.

Han, A. K., and D. Park (1989): "Testing for Structural Change in Panel Data: Application to a Study of U.S. Foreign Trade in Manufacturing Goods," *The Review of Economics and Statistics*, 71(1), 135–42.

Heckman, J. J., R. J. LaLonde, and J. A. Smith (1999): "The economics and econometrics of active labor market programs," in *Handbook of Labor Economics*, ed. by O. Ashenfelter, and D. Card, vol. 3 of *Hanbook of Labor Economics*, chap. 31, pp. 1865–2097. Elsevier.

Heckman, J. J., and S. Navarro (2007): "Dynamic discrete choice and dynamic treatment effects," *Journal of Econometrics*, 136(2), 341–396.

Heckman, J. J., and E. J. Vytlacil (2007): "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6 of *Handbook of Econometrics*, chap. 70. Elsevier.

Horowitz, J. L. (2006): "Testing a Parametric Model against a Nonparametric Alternative with Identification through Instrumental Variables," *Econometrica*, 74(2), pp. 521–538.

——— (2012): "Specification testing in nonparametric instrumental variable estimation," *Journal of Econometrics*, 167(2), 383–396.

Hoxby, C. M. (2000): "The Effects Of Class Size On Student Achievement: New Evidence From Population Variation," *The Quarterly Journal of Economics*, 115(4), 1239–1285.

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 (1997): "Estimating outcome distributions for compliers in instrumental variables models," *Review of Economic Studies*, 64(4), 555–574.

Kalbfleisch, J. D., and R. L. Prentice (2002): *The statistical analysis of failure time data*. John Wiley & Sons, New York.

Kao, C., L. Trapani, and G. Urga (2007): "Modelling and Testing for Structural Changes in Panel Cointegration Models with Common and Idiosyncratic Stochastic Trend," Center for Policy Research Working Papers 92, Center for Policy Research, Maxwell School, Syracuse University.

Karunamuni, R. J., and T. Alberts (2005): "A generalized reflection method of boundary correction in kernel density estimation," *Canadian Journal of Statistics*, 33(4), 497–509.

Katz, L. F., and B. D. Meyer (1990): "The impact of the potential duration of unemployment benefits on the duration of unemployment," *Journal of Public Economics*, 41(1), 45–72.

Kluve, J. (2010): "The effectiveness of European active labor market programs," *Labour Economics*, 17, 904–918.

Lalive, R. (2008): "How do extended benefits affect unemployment duration? A regression discontinuity approach," *Journal of Econometrics*, 142(2), 785–806.

Lalive, R., J. van Ours, and J. Zweimüller (2006): "How changes in financial incentives affect the duration of unemployment," *The Review of Economic Studies*, 73(4), 1009–1038.

Lalive, R., J. Zweimüller, and J. van Ours (2005): "The effect of benefit sanctions on the duration of unemployment," *Journal of the European Economic Association*, 3(6), 1386–1417.

Lancaster, T. (1990): *The Econometric Analysis of Transition Data*. Cambridge University Press.

Lazear, E. P. (2001): "Educational Production," *The Quarterly Journal of Economics*, 116(3), 777–803.

Li, Q., E. Maasoumi, and J. S. Racine (2009): "Educational Production," *Journal of Econometrics*, 148, 186–200.

Mammen, E., J. S. Marron, B. A. Turlach, and M. P. Wand (2001): "A General Projection Framework for Constrained Smoothing," *Statistical Science*, 16(3), pp. 232–248.

Mammen, E., and C. Thomas-Agnan (1999): "Smoothing Splines and Shape Restrictions," *Scandinavian Journal of Statistics*, 26(2), 239–252.

Meyer, B. D. (1996): "What have we learned from the Illinois reemployment bonus experiment?," *Journal of Labour Economics*, 14(1), 26–51.

Moffit, R. (1983): "An economic model of welfare stigma," *The American Economic Review*, 73(5), 1023–1035.

Müller, H. G., and J. L. Wang (1994): "Hazard rate estimation under random with varying kernels and bandwidths," *Biometrics*, 50(1), 61–76.

Nielsen, J. P., and O. B. Linton (1995): "Kernel Estimation in a Nonparametric Marker Dependent Hazard Model," *The Annals of Statistics*, 23(5), 1735–1748.

Pavoni, N., and G. L. Violante (2007): "Optimal Welfare-to-Work Programs," *Review of Economic Studies*, 74(1), 283–318.

Pearl, J. (2000): *Causality. Models, reasoning, and inference*. Cambridge University Press.

Porter, J. (2003): "Estimation in the regression discontinuity model," Mimeo, Department of Economics, University of Wisconsin.

Prieto, A. (2000): "L'impact de la dégressivité des allocations chômage sur le taux de reprise d'emploi," *Revue Éconique*, 51(3), 523–534.

Richardson, K., and G. J. V. den Berg (2001): "The effect of vocational employment training on the individual transition rate from unemployment to work," *Economic Policy Review*, 8(2), 175–214.

Robins, J. M., and A. A. Tsiatis (1991): "Correcting for non-compliance in randomized trials using rank preserving structural failure time models," *Communications in Statistics - Theory and Methods*, 20, 2609–2631.

Rosholm, M., and M. Svarer (2008): "Estimating the threat effect of active labour market programs," *Scandinavian Journal of Economics*, 110(2), 385–401.

Sacerdote, B. (2011): *Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?*vol. 3 of *Handbook of the Economics of Education*, chap. 4, pp. 249–277. Elsevier.

Schneeweis, N., and R. Winter-Ebmer (2007): "Peer effects in Austrian schools," *Empirical Economics*, 32(2), 387–409.

Schunk, D. H. (1991): *Learning theories: an educational perspective*. Merrill, New York.

Sianesi, B. (2004): "An Evaluation of the Swedish System of Active Labor Market Programs in the 1990s," *The Review of Economics and Statistics*, 86(1), 133–155.

Taber, C. R. (2000): "Semiparametric identification and heterogeneity in discrete choice dynamic programming models," *Journal of Econometrics*, 96(2), 201–229.

Trochim, W. (1984): *Research design for program evaluation: the regression-discontinuity approach*, Contemporary evaluation research. Sage Publications.

Tsiatis, G. (1975): "A nonidentifiability aspect of the problem of competing risks," *Proc. Nat. Acad. Sci. 72*, 72(1), 20–22.

Urquiola, M. (2006): "Identifying Class Size Effects in Developing Countries: Evidence from Rural Bolivia," *The Review of Economics and Statistics*, 88(1), 171–177.

Van de Geer, S. (2000): *Empirical processes in M-estimation*. Cambridge University Press.

Van den Berg, G. J. (2001): "Duration models: specification, identification, and multiple durations," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, chap. 55, pp. 3381–3460. Elsevier.

Van den Berg, G. J., A. Bozio, and M. C. Dias (2014): "Policy discontinuity and duration outcomes," *Working Paper*.

Van den Berg, G. J., Kjærsgaard, and M. Rosholm (2012): "To meet or not to meet (your case worker) - that is the question," IZA Discussion Paper 6476, IZA, Bonn, Germany.

Van den Berg, G. J., and B. Van der Klaauw (2010): "Counseling and monitoring of unemployed workers: theory and evidence from a controlled social experiment," *International Economic Review*, 47(3), 895–936.

van der Klaauw, W. (2002): "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach," *International Economic Review*, 43(4), pp. 1249–1287.

Wössmann, L., and M. West (2006): "Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS," *European Economic Review*, 50(3), 695–736.

Declaration

I certify that the thesis I have presented for examination for the Doctor degree of the University of Mannheim is solely my work with exception of the parts and coauthors I have indicated in it. Furthermore, I declare that no sources have been used in the preparation of this thesis other than those referenced in the thesis itself.

Petyo Bonev Mannheim, 8 September 2014

LEBENSLAUF

Petyo Bonev studierte Betriebswirtschaftslehre an der Universität Mannheim (2002 - 2008) mit Spezialisierungsfächern Bankbetriebslehre, Versicherungen, Finanzierung und Statistik. Er absolvierte diverse Praktika, unter anderem bei Heidelberger Druck AG, KPMG AG und The Boston Consulting Group. Seine Promotion im Fach Ökonometrie begann er 2008 an der Graduierten Schule CDSE an der Uni Mannheim. Er verteidigte die Arbeit "Essays in Nonparametric Instrumental Variable Regression" im September 2014 mit der Endnote Magna Cum Laude.