USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

CENTRO INTERNACIONAL DE ESTUDOS
DE DOUTORAMENTO E AVANZADOS
DA USC (CIEDUS)

TESIS DE DOCTORADO

FLEXIBLE MODELS FOR CAUSAL INFERENCE IN MEDICINE AND ECONOMICS

Carlos Matías Hisgen

ESCUELA DE DOCTORADO INTERNACIONAL

PROGRAMA DE DOCTORADO EN ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

SANTIAGO DE COMPOSTELA

2019

UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

DECLARACIÓN DEL AUTOR DE LA TESIS

"Flexible Models for Causal Inference in Medicine and Economics"

D. Carlos Matías Hisgen

Presento mi tesis, siguiendo el procedimiento adecuado al Reglamento, y declaro que:

1) *La tesis abarca los resultados de la elaboración de mi trabajo.*
2) *En su caso, en la tesis se hace referencia a las colaboraciones que tuvo este trabajo.*
3) *La tesis es la versión definitiva presentada para su defensa y coincide con la versión enviada en formato electrónico.*
4) *Confirmo que la tesis no incurre en ningún tipo de plagio de otros autores ni de trabajos presentados por mí para la obtención de otros títulos.*

En Ciudad de Resistencia, Argentina, 07 de julio de 2019

Fdo. Carlos Matías Hisgen

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

AUTORIZACIÓN DEL DIRECTOR / TUTOR DE LA TESIS

"Flexible Models for Causal Inference in Medicine and Economics"

Dña. Carmen María Cadarso Suárez
D. Daniel Gustavo Miles Touya

INFORMA/N:

Que la presente tesis, corresponde con el trabajo realizado por D. ***Carlos Matías Hisgen****, bajo nuestra dirección, y autorizamos su presentación, considerando que reúne l os requisitos exigidos en el Reglamento de Estudios de Doctorado de la USC, y que como directores de ésta no incurre en las causas de abstención establecidas en Ley 40/2015.*

En Santiago de Compostela, 10 de Julio de 2019

Fdo. Carmen María Cadarso Suarez

Fdo. Daniel Gustavo Miles Touya

39452691D

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

DISSERTATION

FLEXIBLE MODELS FOR CAUSAL INFERENCE IN MEDICINE AND ECONOMICS

Author:
Carlos Matías HISGEN

Advisors:
Carmen María CADARSO SUÁREZ
Daniel Gustavo MILES TOUYA

DEPARTMENT OF STATISTICS,
MATHEMATICAL ANALYSIS AND
OPTIMIZATION

July 3, 2019

Acknowledgements

During the time this thesis was written several people and actors intervened to make possible its realization. In first place I am indebted to my supervisor Carmen Cadarzo Suarez, who first told me about the USC's Doctoral Program in Statistics and helped me during my early years as a foreign student. I also thank to my co-supervisor Daniel Miles Touya, who permanently guided and inspired me to find the adequate research questions.

I'd like to acknowledge Emad Abu-assi, Manuel Wiesenfarth and Thomas Kneib who collaborated with my research and all the program's professors, specially to Alberto Rodriguez Casal and Javier Roca Pardiñas for they support and motivation. I am also grateful to other colleagues and friends: Mikel Álvarez Mozos, Isabel Martínez Silva, Ivan Folgueira Pérez, María Xosé Rodríguez Álvarez and Juan Jose Santos. Finally, I want to thank my whole family and my couple Sonia for their support.

Fundings: I am very grateful for the financial support during my stay in Spain, which was provided by two institutions, Agencia Española de Cooperación Internacional para el Desarrollo (Programa de Becas MAEC-AECID, 2009-2010 and 2010-2011) and Universidad Nacional del Nordeste (UNNE).

FLEXIBLE MODELS FOR CAUSAL INFERENCE IN MEDICINE AND ECONOMICS

Carlos Matías HISGEN

Abstract

The aim of the present work is the study of empirical aspects of a flexible regression procedure designed to perform causal inference, known as the Nonparametric Triangular Simultaneous Equations Model. This procedure helps to mitigate a problem that arise when the model regressors do not fulfill the exogeneity assumption. The main contributions emerge from two empirical applications, in Medicine and Economics, and a new bayesian estimator which is evaluated by Monte Carlo simulation. The first application involves an implementation of the triangular simultaneous equations model to assess the effects of a treatment, defined as *time delay to catheterization*, on the outcome, defined in terms of survival and cardiac health, for patients with non ST-segment elevation Myocardial Infraction. The main methodological contribution consists on modeling the treatment as a continuous variable, instead of using a dichotomous variable indicating early versus late intervention, and using a flexible Generalized Additive Model for estimation and inference. The second application pursue an estimation of the class size's effect on schooling achievement (measured by Literature's test-scores), for students from sixth grades of the primary school in Uruguay. Main innovations consist on both, implementation of a flexible additive model that enables us to take into account nonlinear effects of control variables, and perform an adequate trimming of outlier observations, which are usually ignored in similar applications. The bias caused by these outliers is illustrated by a Monte Carlo simulation exercise. Finally, the simulation study addressees the problem of weak identification in the nonparametric instrumental variable framework. In particular, it assess the performance of two alternative non-parametric estimators of the Triangular Simultaneous Equations Model when weak instruments are present. Two estimators are compared, the Two Stage Generalized Additive Model (2SGAM) and a new Bayesian Nonparametric Instrumental Variables (BNIV) estimator. Simulation results support the advantages of BNIV over 2SGAM when instruments are weak. Specifically, when the concentration parameter ranges between 10 and 16, BNIV outperform 2SGAM in terms of variance. The mentioned efficiency advantage of BNIV does not imply an increment in bias.

Keywords: nonparametric, endogeneity, triangular simultaneous equations, instrumental variables.

MODELOS FLEXIBLES PARA INFERENCIA CAUSAL EN MEDICINA Y ECONOMIA

Carlos Matías HISGEN

Resumen

El objeto del presente trabajo consiste en el estudio de aspectos empíricos de un procedimiento de regresión diseñado para realizar inferencia causal, conocido como Modelo No-paramétrico de Ecuaciones Simultáneas Triangulares. Este procedimiento ayuda a mitigar un problema que surge cuando los regresores del modelo no cumplen con el supuesto de exogeneidad. Las principales contribuciones provienen de dos aplicaciones empíricas, en Medicina y Economía, y un nuevo estimador bayesiano el cual es evaluado mediante simulación Monte Carlo. La primer aplicación involucra la implementación de modelos de ecuaciones triangulares simultáneas para evaluar los efectos de un tratamiento, definido como *tiempo de retraso hasta la cateterización*, sobre una medida de resultado, definida en términos de supervivencia y de salud cardíaca, para pacientes con Infarto Agudo de Miocardio sin elevación del segmento ST. La principal contribución metodológica consiste en modelar el tratamiento como una variable continua, en vez de representarla como una variable dicotómica indicando intervención temprana versus tardía, y usando un Modelo Aditivo Generalizado. La segunda aplicación conlleva la estimación del efecto del número de alumnos en la clase sobre el rendimiento escolar (medido mediante calificaciones en evaluaciones de literatura), para estudiantes del sexto grado de la escuela primaria en Uruguay. Las principales innovaciones consisten en la implementación de un modelo aditivo flexible, que permite considerar efectos no lineales para las variables de control, y en la adecuada exclusión de observaciones atípicas, las cuales son ignoradas frecuentemente en aplicaciones similares. El sesgo causado por estas observaciones atípicas es ilustrado mediante un ejercicio de simulación Monte Carlo. Finalmente, el estudio de simulación aborda el problema de identificación débil en el marco de la regresión no-paramétrica con variables instrumentales. En particular, se evalúa el desempeño de dos estimadores no-paramétricos alternativos, para el Modelo de Ecuaciones Simultáneas Triangulares, cuando los instrumentos son débiles. Dos estimadores son comparados, el Modelo Aditivo Generalizado en Dos Etapas (2SGAM) y un nuevo estimador Bayesiano No-paramétrico con Variables Instrumentales (BNIV). Los resultados avalan las ventajas del estimador BNIV por sobre el 2SAM cuando los instrumentos son débiles. Específicamente, cuando el parámetro de concentración se encuentra entre 10 y 16, el BNIV aventaja al 2SGAM en términos de varianza. La mencionada ventaja en eficiencia del BNIV no implica un incremento relativo en términos de sesgo.

Palabras clave: no-paramétrico, endogeneidad, ecuaciones simultáneas triangulares, variables instrumentales.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

A main concern in empirical research is to uncover causal relationships. More precisely, whether a particular intervention or treatment causes, explains or motivates a particular effect or outcome. For example, does a reduction in the size of a school class increase test scores? Or, does unemployment training programs affect the length of unemployment spells or ex-post program income?

In the presence of randomized experiments, with random assignment of the treatment among individuals or study units, it is relatively simple to derive causal conclusions. Basically, in this case we compare the average outcome for individuals in the treated group against the average outcome in the non-treated group. The randomization mechanism tends to balance observable and unobservable characteristics making groups comparable.

In contrast, identifying causal relationships in observational studies, where the mechanism that assigns individuals to different treatment states is unknown or not random (i.e. the analysis is performed using non-experimental or observational data) is not so simple. In this case, individuals in both groups can be systematically different in terms of unobservable characteristics, confounding the causal effects of the treatment. For example, an individual's decision to participate in an unemployment training program may depend on the outcome of the training program. Therefore, in this case the statistical model should incorporate explicitly how individuals decide to participate or not in the program.

This work is intended to extend the empirical knowledge and possibilities of a flexible regression model designed to perform causal inference in empirical sciences when the data comes from an observational process. This model, technically known as the Nonparametric Triangular Simultaneous Equations Model, helps mitigate a problem that arises when the model regressors or covariates do not satisfy the condition known as *exogeneity assumption*, which establishes that the model's random component must be *mean independent* from all the model regressors.

This chapter introduces core concepts and methods which are extensively used in subsequent chapters and, in its final section, presents the structural ordering of the work.

It is remarked that definitions presented here are general, leaving the details to be specified in each chapter. Moreover, all chapters were written to be self-contained.

The following two sections define some general notions about *causal inference* in regression analysis with non-experimental data, the concept of *endogeneity* and the *Control Function Approach* to instrumental variables regression.

1.1 Causal inference in regression analysis with observational data

Regression analysis with non-experimental (or observational) data is often used in social or life sciences to infer the existence of a causal relation between a *treatment* variable x and a *response* variable y (measuring the outcome affected by the treatment x). The presence (or absence) of a simple statistical relationship among those variables is not a sufficient nor a necessary condition to claim the presence (or absence) of a causal relationship. This is so because the measured values of both treatment and response variables are generated by an uncontrolled experiment (e.g. a natural or social process) which is in principle unknown by the researcher (for discussions of these ideas in sociology, natural sciences and economics see Winship and Morgan, 1999, Rosenbaum, 1984, Rosenbaum, 2002 and Heckman, 2008).

Therefore, prior specification of a theoretical or structural model, establishing a causal link and a causal direction between the treatment and the response, is required. Moreover, such a model must take into account that treatment x is not usually randomly assigned to the population units and, as result, it may be related to other factors, say $\mathbf{z}$, which systematically affect the response y in addition to x.

In accordance to the previous discussion and the family of models employed in this thesis, consider the following mean regression model with additive error (1.1), assumed to represent the true population model or Data Generating Process (DGP).

$$y = f(x, \mathbf{z}) + \varepsilon, \tag{1.1}$$

where x and y are continuous variables, $\mathbf{z}$ is a vector containing other observable or measurable factors (which can be continuous and/or discrete variables related to x) that *systematically* affect outcome y, and ε is an additive error component representing the effects of unobserved or unmeasured factors which affect response y. Note that x, $\mathbf{z}$ and ε represent all possible factors determining y (i.e. once the values of those factors are established then the level of the outcome y is completely determined).

Assuming that ε is mean independent of x and $\mathbf{z}$ is central for identification of a causal relationship between x and y. This condition can be statistically expressed by moment restriction

(1.2).

$$E(\varepsilon|x,\mathbf{z}) = 0 \tag{1.2}$$

Restriction (1.2) is usually known as the *exogeneity assumption*, and it implies that the error component of the response y is mean independent from the treatment x once conditioning on factors $\mathbf{z}$. Note that condition (1.2) is accomplished if treatment x is randomly assigned to population units implying that $E(\varepsilon|x) = 0$, but as mentioned above this random assignment is not possible when observational data is used.

Taking conditional expectation over (1.1), and given (1.2), it is possible to identify the mean regression function (1.3),

$$E(y|x,\mathbf{z}) = f(x,\mathbf{z}) + E(\varepsilon|x,\mathbf{z}) = f(x,\mathbf{z}), \tag{1.3}$$

which is the expected value of y conditional on x and $\mathbf{z}$.

Identification of the regression function (1.3), causally linking y to x, enables the researcher to identify the *marginal effect* of treatment x, defined as the partial derivative (1.4)

$$\frac{\partial E(y|x,\mathbf{z})}{\partial x} = \frac{\partial f(x,\mathbf{z})}{\partial x}. \tag{1.4}$$

Estimating the marginal treatment effect (1.4) is usually the main goal of causal inference analysis based on regression methods. It can be interpreted as: *the marginal change in the expected value of response y caused by a marginal change in treatment x, when the additional factors in z remain constant.* As we noted earlier, this causal interpretation is derived from hypothetical causality channels, based on prior theoretical models, which must be incorporated in DGP (1.1).

Therefore, from a theoretical point of view, the non-experimental nature of the data requires obtaining the DGP (1.1) from a theoretical model establishing the causal structure that connects the involved variables. In concrete, satisfying the exogeneity assumption (1.2) requires the DGP design avoids the effects of the following scenarios (usually present in observational studies):

- S.1. The existence of a mechanism that simultaneously determines the values of both, the response y and the treatment x. This possibility, called the *simultaneity problem*, is a common issue in econometric applications such as estimation of demand and supply functions (see Haavelmo, 1943 for an early analysis).

- S.2. The presence of a *self-selection problem*, which arises when the individuals under analysis can choose the level of treatment x taking into account its expected effect over outcome y (seminal works identifying this issue are Gronau, 1973 and Heckman, 1974;

see Heckman, 1990 and Heckman and Vytlacil, 2007 for more recent presentations). For example, a person deciding whether or not to enroll in a graduate study program by assessing the impact of such decision on her expected future income.[1]

- S.3. Presence of the *reverse causality problem*. Such a problem arises when not only the treatment x has an effect on response y but also y has an impact over x. Situations like that can be generated by the *simultaneity problem* previously mentioned or by dynamic interrelationships between x and y due to future expectations. For example, the actual provision of urban police in time t (i.e. the treatment x_t) affects the level of actual urban crime (i.e the response y_t), but it is possible that the expected future value of the level of urban crime (y_{t+1}) affects the actual provision of urban police (x_t).

Accounting for these potential scenarios (S.1 to S.3), often requires constructing formal models that explicitly define the *assignment mechanism* of treatment x, specially in behavioral science as Economics and Sociology (see for example Heckman, 2008 and Heckman, 2005).

On the other hand, from an empirical point of view involving an application to real data, achievement of exogeneity assumption (1.2), in the context of DGP (1.1), requires the simultaneously accomplishment of the following conditions:

- C.1. All relevant factors in $\mathbf{z}$ must be measured and included in the model as regressors. These factors are usually known as control variables or co-variables. This condition is violated when some elements of $\mathbf{z}$ are ignored and excluded from the analysis. For example, some factors in $\mathbf{z}$ can be unobservable for the researcher, therefore they cannot be measured and included in the empirical model. This situation is known as the *omitted variables* or *unobserved confounding* problem.

- C.2. Specification of the regression model must be *close enough* to the true DGP (1.1). This mainly involves specifying the functional form linking regressors x and $\mathbf{z}$ with the outcome y (i.e. defining the functional form of $f(\cdot)$ in (1.1)) and, if necessary for special kind of DGP, specifying the probability distribution of the random component ε. Satisfaction of this condition fails when functional forms involved in the DGP are incorrectly specified, for example if the effect of a regressor is defined as a linear function when the true effect is nonlinear. This situation is known as the *model misspecification problem*.[2]

- C.3. All relevant variables, y, x and those in $\mathbf{z}$, must be measured without error. Failure of this requirement originates a *measurement error problem* (see Wooldridge, 2010 for an exposition of the more frequent cases in econometrics).

[1] The *self-selection problem* can be cast into a more general category called *sample selection problem*, under which the observable samples are not representative of the population under study. Therefore, the sample selection problem can be generated by different sources, as is the censoring or truncation of the dependent variable (see Maddala, 1986 for a survey in the linear model context.)

[2] A third aspect to be defined relates the functional form through which the random term ε affects the response y, that we established as an additive function in DGP (1.1).

If one or more empirical conditions, C.1 to C.3, are not satisfied and/or theoretical scenarios, S.1 to S.3, are not correctly handled, then the exogeneity assumption (1.2) is violated. Under such a situation, known as the *endogeneity problem* in the econometric literature, regression function (1.3) is replaced by (1.5).

$$E(y|x,\mathbf{z}) = f(x,\mathbf{z}) + E(\varepsilon|x,\mathbf{z}). \tag{1.5}$$

From (1.5) it can be seen that the regression function of interest, $f(x,\mathbf{z})$, cannot be identified because expected value $E(\varepsilon|x,\mathbf{z})$ is a non-constant function of x and/or $\mathbf{z}$. It is important to note that such an expected value cannot be estimated because error term ε is unobserved. In other words, under these circumstances, the usual parametric and non-parametric estimators for the regression of y over x and $\mathbf{z}$ will be inconsistent.

For example, assuming $\varepsilon = \zeta + \varepsilon$, model (1.1) can be rewritten as (1.6)

$$y = f(x,\mathbf{z}) + \varepsilon = f(x,\mathbf{z}) + \zeta + \varepsilon, \tag{1.6}$$

where $E(\varepsilon|x,\mathbf{z}) = 0$ and each individual in the population may choose an optimal level of treatment x according to the level of unobserved factor ζ, so that expectation $E(\zeta|x,\mathbf{z}) = E(\zeta|x)$ systematically varies with x. In this case, the conditional expectation of y is given by (1.7)

$$E(y|x,\mathbf{z}) = f(x,\mathbf{z}) + E(\varepsilon|x,\mathbf{z}) = f(x,\mathbf{z}) + E(\zeta|x). \tag{1.7}$$

Since factor ζ is known by the individual but is not observable for the researcher, the function $f(x,\mathbf{z})$ is not straightforwardly identified. Therefore, in this kind of self-selection problem, an explicit behavioral model of treatment selection must be considered.

A typical applied example of model (1.6) under condition (1.7) consist on explaining *individual earnings* of workers in terms of the treatment *years of formal education* and other control variables. A potential self-selection problem will be present in this example if there exist an unobserved factor as the *innate ability* of workers which determines both the *individual earnings* and the *years of formal education*.

Some intuitive illustration of above mentioned self-selection problem is given by the following path diagram

$$\begin{tikzcd}
x = years\,of\,education \arrow[r] & y = individual\,earnings \\
\zeta = innate\,ability \arrow[u] \arrow[ur] &
\end{tikzcd}$$

where years of education and innate ability have a direct effect on earnings, but additionally innate ability possess a direct effect on education. This additional effect imply that innate ability has an indirect effect on earnings through its association with years of education, producing the endogeneity problem.

1.2 Endogeneity, instrumental variables and the control function approach

To overcome the endogeneity problem described in the previous section several methodologies have been developed. One of the pioneer methods in the field is the Instrumental Variables Regression (IVR), first developed in the context of the Linear Regression Model.

The IVR was originaly applied with the parametric Linear Regression Model to deal with the simultaneity problem in the context of demand and supply functions estimation (see for example the seminal application of IV regression in Wright, 1928). Nowadays, applying IVR with the linear model represents a standard resource in the econometric toolkit, which can be estimated by several methods as Two-Stage Least Squares, Restricted Maximum Likelihood and Generalized Methods of Moments (excellent technical and applied presentations can be found in Bowden and Turkington, 1990, Angrist and Pischke, 2009 and Wooldridge, 2010).[3]

In general, the instrumental variables (IV) method relies on the existence of at least one additional variable (i.e. the instrument) for each endogenous regressor in the model. Intuitively, this instrument must be correlated to its corresponding endogenous regressor and uncorrelatd to any other variable factor in the model. For DGP (1.1), and assuming that the only endogenous variable is the treatment x (i.e. $E(\varepsilon|x,\mathbf{z}) = E(\varepsilon|x) = \psi(x)$, where $\psi(.)$ is a general function of x), it is enough to have one instrument that we set as w. Then, w is the instrument for the endogenous treatment x and may consist of a continuous or a discrete variable (note that it is possible to have more than one instrument, but for simplicity of exposition we use only one).

To accomplish identification of the regression function of interest, the instrument w must satisfy some specific conditions:

- ID.1 The instrument w must have explanatory power over the treatment x. This alludes to the degree of conditional or partial association between the treatment x and the instrument w (given the additional controls $\mathbf{z}$), which is linked to the notion of *strong* instruments as opposed to *weak* instruments. [4]

[3] There have been additional developments of IV estimators in a Bayesian framework (see for example Dreze and Richard, 1983 and Kleibergen and Van Dijk, 1998). We limit our attention to the frequentist view, at least in this introduction, since the basic concepts we want to review are the same for both, frequentist and bayesian approaches.

[4] A formal definition of weak instruments and discussion of their effects are provided in Chapter 4.

For example, assuming that the true regression model explaining x is linear and that there is a unique control variable z, as in (1.8),

$$x = E(x|w,z) + u = \alpha_0 + \alpha_1 w + \alpha_2 z + u, \tag{1.8}$$

then, condition ID.1 is satisfied if $\alpha_1 \neq 0$. This condition can be tested applying a standard test of hypothesis with the null $H_0 : \alpha_1 = 0$.

This example can be straightforwardly extended when w (or z) has a nonlinear effect.

- ID.2 The instrument must be unrelated to the unobserved factors ε which determine the outcome variable y in (1.1), in the way established by moment restriction (1.9).

$$E(\varepsilon|w,\mathbf{z}) = 0. \tag{1.9}$$

This condition ensures that w is not a relevant explanatory variable for y in DGP (1.1). This means, jointly with condition ID.1, that instrument w only affects the outcome y through its effect over the endogenous treatment x.

The following path diagram presents a visual illustration of the IV assumptions,

$$\begin{tikzcd}
w \arrow[r] & x \arrow[r] & y \\
& \varepsilon \arrow[u] \arrow[ur] &
\end{tikzcd}$$

where instrument w has a direct effect on x, an indirect effect on y and is not associated with unobserved component ε.

The goal of identification assumptions ID.1 and ID.2 is to establish an exogenous source of variation in treatment x, via the instrument w. Using the variability of w in this manner is equivalent to obtain an assignment of the treatment x (over the population) that is not influenced by ε.

It is important to note that assumptions ID.1 and ID.2 must be justified or derived from a prior theoretical model, specially assumption ID.2 which cannot be tested from available data (ε is unobservable for the researcher).

1.2.1 Instrumental variables in the Simple Linear Model

To better understand how an IV estimator works, it's useful to analyze the following two equations system (1.10), based on simple linear regression models and assuming a random sample of observations of size n.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad x_i = \alpha_0 + \alpha_1 w_i + u_i, \tag{1.10}$$

where $\{y_i, x_i, w_i\}_{i=1}^{n}$ are n realizations of the outcome variable, the treatment, and the instrument, respectively, and β_1 is the coefficient we want to estimate consistently.

If treatment x were an exogenous variable in model (1.10) then the Ordinary Least Squares (OLS) estimator for parameter β_1 would be consistent. This OLS estimator, obtained from adjusting the first equation in (1.10), can be expressed as in (1.11),

$$\hat{\beta}_1^{OLS} = \frac{Cov(y_i, x_i)}{Var(x_i)}, \tag{1.11}$$

where $Cov(y_i, x_i)$ and $Var(x_i)$ are the sample covariance between y and x and the sample variance of x, respectively.

On the other hand, if x were an endogenous regressor, then the OLS estimator for β_1 would be inconsistent. However, under assumptions ID.1 and ID.2, the instrumental variable estimator (1.12) would be consistent.

$$\hat{\beta}_1^{IV} = \frac{Cov(y_i, w_i)}{Cov(x_i, w_i)}. \tag{1.12}$$

From the definition (1.12) it can be seen that the IV estimator uses only the part of the variation in x that is correlated with w, leaving aside the part of x which may be correlated to unobserved factors in ε (remember from ID.2 that w is unrelated with ε).

An usual simple procedure to get $\hat{\beta}_1^{IV}$ is the so called Two-Stage Least Squares (2SLS) which can be defined in two steps:

- First Stage: use OLS to estimate parameters α_0 and α_1 of the second equation in (1.10), and obtain the fitted values $\hat{x}_i = \hat{\alpha}_0 + \hat{\alpha}_1 w_i$.
- Second Stage: regress outcome y over the fitted values $\hat{x}$ (including an intercept). In other words, adjust by OLS the first equation in (1.10) but replacing x_i by its fitted values $\hat{x}_i$. The coefficient of $\hat{x}_i$ in this regression is equivalent to the IV estimator (1.12).

There is an alternative two-stage procedure that yields the same IV estimator (1.12), proposed by Hausman, 1978 and Hausman, 1983 in order to test the exogeneity status of a potential endogenous regressor. The procedure is the following:

- First Stage: use OLS to estimates parameters α_0 and α_1 of the second equation in (1.10), and obtain the residuals $\hat{u}_i = x_i - \hat{x}_i = x_i - (\hat{\alpha}_0 + \hat{\alpha}_1 w_i)$.
- Second Stage: regress outcome y over treatment x and residuals $\hat{u}$ (including an intercept). In other words, adjust by OLS the first equation in (1.10) but including the first-stage residuals as an additional control variable.

As in the 2SLS case, this second procedure uses the first stage to split the variability of the endogenous regressor x in two parts. The first part ($\hat{x}$) which is generated by instrument w, is considered as the strictly *exogenous* portion. And the second part $\hat{u}$, that is orthogonal to z, is qualified as the potentially endogenous portion. Then, in the second stage estimation, this IV method explicitly controls for the endogenous part of x.

The later two-stage procedure, sometimes called Two-Stage Residual Inclusion, is the antecedent of the so called Control Function Approach to instrumental variables, which we describe in the next subsection.

1.2.2 Instrumental variables in the Nonparametric Model

When the DGP of interest involves a general (i.e. nonparametric) regression function such as (1.1) and the regressor of interest is endogenous, it is necessary to rely on a nonparametric IV estimator.

Earlier works attempting to realize such an extension (i.e. relying on moment restriction (1.9)) are Ai and Chen, 2003, Newey and Powell, 2003 and Hall and Horowitz, 2005. These and subsequent works derived usual statistical properties for the new estimators but all of them share the problem related to the lack of an optimal rule for choosing the smoothing parameter (or regularization parameter), which is crucial in any nonparametric curve estimation. Recently, Horowitz, 2014 proposes for the first time a nonparametric IV series estimator accompanied with a theoretically justified method for choosing the smoothing parameter. Mentioned developments, based on identification condition (1.9), are framed into the so called *Regularization Approach* to nonparametric IV estimation.

Parallel to the developments within the Regularization Approach, another formulation of the nonparametric IV regression, called the *Control Function Approach* (CFA), was first proposed by Newey, Powell, and Vella, 1999 and extended by Pinkse, 2000, Su and Ullah, 2008 and Marra and Radice, 2011. This alternative approximation is neither more nor less general than the Regularization Approach, because it relies on a different set of identification assumptions.

The CFA approach is based in a Triangular Nonparametric Simultaneous Equations Model (Newey, Powell, and Vella, 1999) defined by (1.13), (1.14) and (1.15):

$$y = f(x,\mathbf{z}) + \varepsilon, \tag{1.13}$$

$$x = g(w,\mathbf{z}) + u, \tag{1.14}$$

$$E(u|w,\mathbf{z}) = 0, \qquad E(\varepsilon|u,w,\mathbf{z}) = E(\varepsilon|u), \tag{1.15}$$

where (1.13) is the same equation than DGP (1.1), $g(\cdot)$ is an unknown function with $g(w,\mathbf{z}) = E(x|w,\mathbf{z})$, and u is an usual error component. Again, we assume the availability of only one instrument w for endogenous treatment x.

To obtain identification, the CFA maintains assumption ID.1 and replaces moment restriction (1.9), defined by assumption ID.2, with the conditions in (1.15). The first moment restriction in (1.15) establishes that the error u is mean independent of instrument w and control variables $\mathbf{z}$. The second restriction in (1.15) ensures that, after conditioning on u, the error term ε is mean independent of instrument w and controls $\mathbf{z}$.

Following Newey, Powell, and Vella, 1999, we combine equations (1.13), (1.14) with conditions in (1.15), to obtain the following expected value for y:

$$\begin{aligned} E(y|x,w,\mathbf{z}) &= f(x,\mathbf{z}) + E(\varepsilon|x,w,\mathbf{z}) = f(x,\mathbf{z}) + E(\varepsilon|u,w,\mathbf{z}) \\ &= f(x,\mathbf{z}) + E(\varepsilon|u) = f(x,\mathbf{z}) + f_u(u). \end{aligned} \tag{1.16}$$

Therefore, identification of the regression function of interest, $f(x,\mathbf{z})$, requires taking into consideration an additional term, i.e. the *control function* $E(\varepsilon|u) = f_u(u)$.

Conditional expectation (1.16) explicitly controls for the variability in the endogenous regressor x which is related to the error term ε (i.e. it controls for the variability of error term u). This enables to identify function $f(\cdot)$ and marginal effect (1.4) using only the variability of x that is unrelated with error term ε.

To perform estimation in the CFA's framework, it is possible to decompose the estimation problem into two sequential stages, using a procedure similar to that used by the Two-Stage Residual Inclusion method. Given a random sample of observations of the relevant variables, $\{y_i, x_i, \mathbf{z}_i, w_i\}_{i=1}^{n}$, the two stages can be described as follows:

- First Stage: function $g(w,\mathbf{z})$ in (1.14) can be estimated by a standard nonparametric regression estimator, obtaining

  $$x_i = \hat{g}(w_i,\mathbf{z}_i) + \hat{u}_i,$$

  then the corresponding residuals $\hat{u}_i = x - \hat{g}(w_i,\mathbf{z}_i)$ can be computed. These residuals estimate the true errors u consistently.

- Second Stage: a non-parametric estimator can be used again to estimate functions $f(\cdot)$ and $f_u(\cdot)$ in equation (1.16), using first-stage's residuals $\hat{u}_i$ as a regressor instead of unobserved errors u_i:

  $$\hat{y}_i = \hat{f}(x_i,\mathbf{z}_i) + \hat{f}_u(\hat{u}_i).$$

Because $\hat{u}$ is the estimated component of treatment x that co-varies with unobserved error ε (generating the endogeneity problem), estimating control function $\hat{f}_u(\hat{u})$ makes it possible to isolate the treatment effect generated by the exogenous portion of variability in x.

1.2.3 The Control Function Approach in the Additive and Generalized Additive Models

Through the rest of the thesis we use the CFA methodology to estimate several regression functions similar to the one described in DGP (1.1), but with an additive structure including both nonparametric and parametric terms. As is known in nonparametric regression theory and observed in Newey, Powell, and Vella, 1999, such additive or semiparametric model helps to avoid the curse of dimensionality which emerges in nonparametric estimation when there is a large number of regressors.

For example, given a DGP similar to the system (1.13)-(1.14) but setting w as a continuous variable and assuming $\mathbf{z} = (z_1, z_2)$ (where z_1 is a binary variable and z_2 is a continuous variable), the additive regression system is specified as

$$y = \beta_0 + \beta_1 z_1 + f_1(x) + f_2(z_2) + \varepsilon, \tag{1.17}$$

$$x = \alpha_0 + \alpha_1 z_1 + g_1(w) + g_2(z_2) + u, \tag{1.18}$$

where each regression equation belongs to the semiparametric Additive Model framework (Hastie and Tibshirani, 1986).

Moreover, system (1.17)-(1.18) can be extended to the case in which one or both equations are specified as Generalized Additive Models (GAM) (Hastie and Tibshirani, 1986). This extension, proposed by Marra and Radice, 2011, is illustrated by system (1.19)-(1.20)

$$y = l_2(\beta_0 + \beta_1 z_1 + f_1(x) + f_2(z_2)) + \varepsilon, \tag{1.19}$$

$$x = l_1(\alpha_0 + \alpha_1 z_1 + g_1(w) + g_2(z_2)) + u, \tag{1.20}$$

where $l_r(\cdot)$, $r = 1,2$, is defined as $l_r = \iota_r^{-1}$, and ι_r^{-1} is known as the *link function*.

The link function is smooth and monotonic and it's useful because it imposes boundaries to the response variables values (see McCullagh and Nelder, 1989 for a definition of the link function in the context of Generalized Linear Models). Typical examples are the Probit and Logit link functions, which are used when the response is a binary or Bernoulli random variable. The Probit link consist in the Accumulated Density Function (ADF) of the Standard Normal Distribution, and the Logit link is the ADF of a Standard Logistic Distribution. One of the empirical applications presented in Chapter 2 involves estimation of a GAM with a Probit link function in the main regression equation.

As in the previous cases, and as established in Marra and Radice, 2011, this simultaneous equations system can be estimated by a two steps procedure as follows:

- First Stage: the Generalized Additive Model (1.20) can be estimated by a consistent estimator,

  $$x_i = l_1(\hat{\alpha}_0 + \hat{\alpha}_1 z_{1i} + \hat{g}_1(w_i) + \hat{g}_2(z_{2i})) + \hat{u}_i,$$

  and the residuals $\hat{u}_i = x - l_1(\hat{\alpha}_0 + \hat{\alpha}_1 z_{1i} + \hat{g}_1(w_i) + \hat{g}_2(z_{2i}))$ can be obtained.

- Second Stage: a consistent estimator can be applied to estimate the regression function in (1.19), including first-stage's residuals $\hat{u}_i$ as an additional regressor instead of unobserved errors u_i:

  $$\hat{y}_i = l_2(\hat{\beta}_0 + \hat{\beta}_1 z_{1i} + \hat{f}_1(x_i) + \hat{f}_2(z_{2i}) + \hat{f}_{\hat{u}}(\hat{u}_i)).$$

The estimator of the function of interest, $\hat{f}_1(x_i)$, that emerges from this two-stages procedure is consistent under the assumption that instrument w is independent of error component ε (Marra and Radice, 2011), which is a more restrictive assumption than the one implied by (1.15).

The flexible models involved in the two-stages procedures described above can be estimated by alternative methods. Local Polynomial (kernel) regression methods can be used as in Su and

Ullah, 2008. On the other hand, Newey, Powell, and Vella, 1999, Pinkse, 2000 and Marra and Radice, 2011 employ series expansion estimators.

A series estimator represents the (unknown) functional form of the principal effect, for each continuous regressor, as an infinite series of *approximating functions*, also known as *basis functions*.

Following Marra and Radice, 2011, we use series estimators based on spline basis functions. There are alternative types of spline basis (e.g. B-splines, Thin Plate splines and cubic splines). Estimation under this spline representation can be implemented by Penalized Regression Spline Approach, introduced by Eilers and Marx, 1996 (see Wood, 2006a for an extended exposition). This approach uses a roughness penalty during the model-fitting process to avoids the problem of overfitting. More details of the estimation procedure with Penalized Splines are presented in Chapter 4.

In the following chapters we use different types of spline basis functions and estimation algorithms which we select for each case based on practical considerations.

1.3 Thesis Structure

In this chapter we have introduced core concepts regarding causal inference in regression models, the endogeneity problem and the instrumental variables estimation methodology. The next four chapters present advances in the empirical application of flexible additive models when the treatment variable is endogenous and introduces new insights on the weak identification problem in nonparametric regression.

Chapter 2 consists in an empirical application of the CFA methodology to assess the effects of a treatment, defined as *time delay to catheterization*, on the outcome, defined in terms of *survival and cardiac health*, for patients with non ST-segment elevation Myocardial Infraction. The main medical interest consists in identifying the optimal timing to intervention (i.e. catheterization) in patients with high risk. In such setting the treatment variable is expected to be endogenous; accordingly, standard regression methods are inadequate and an instrumental variable estimator is applied. As in previous studies in the literature, we exploit the exogenous variability in the treatment induced by the fact that patients who arrive at the hospital over the weekends are more prone to experience catheterization delays. The main methodological contribution consists in modeling the treatment as a continuous variable (i.e. *continuous time*), instead of using a dichotomous variable indicating *early* versus *late* intervention, and using a flexible Generalized Additive Model for estimation and inference. This innovation enables us to estimate a nonlinear treatment effect and to evaluate its magnitude over the entire range of the treatment regressor. The estimation results, which support the existence of a significant treatment effect, suggest that usual parametric models can produce a downward estimation bias in the average effect.

In Chapter 3, we pursue an estimation of the effect of *class size* on schooling achievement (measured by Literature's test-scores), for sixth grade students of the primary school in Uruguay. The main obstacles to overcome in this type of application are the endogenous status of the treatment variable (i.e. the class size) and the clustered structure of the data at the school level. To construct an instrumental variable for class size, we take advantage of regulation laws in Uruguay, that establish in 40 students the upper limit for the class size. The main innovations are the application of a flexible additive model, that enables us to take into account nonlinear effects of control variables, and the implementation of a flexible bootstrap methodology for confidence interval construction, in the presence of clustered observations. Additionally, an adequate trimming of outlier observations is performed, which avoids bias in the first stage of the estimation procedure. In this line, a simulation exercise is presented illustrating the bias induced by outlier observations. Overall results provide support for the usefulness of the proposed innovations for the identification of the class size effect.

Chapter 4 addresses the problem of *weak identification* or *weak instruments* in the nonparametric instrumental variable framework. In concrete, it presents an evaluation and comparison of two alternative methods, the frequentist Two-Stage Generalized Additive Model and a new Bayesian Nonparametric Instrumental Variables model. The bayesian method, proposed by Wiesenfarth et al., 2014 and derived from part of the work in this thesis, seems to present advantages in weak instruments scenarios. The weak instruments problem, which represents an important issue for the applied researcher, was largely neglected in the nonparametric literature. One important reason for this neglect has been the difficulty in the development of a flexible instrumental variable estimator, with a suitable method for smoothing parameter selection and a valid inference procedure. This difficulty was solved by the two alternative models that we compare. In particular, the bayesian model allows us to estimate the smoothing parameter from data taking into account the simultaneity nature of the triangular equations system involved. The simulation results imply an advantage of the bayesian method over the frequentist approach, in terms of variance reduction, when instruments are close to being weak (in terms of the parametric literature on weak identification).

Finally, for each of the described chapters, some complementary information (including estimation and inference details) and selected R code, are presented in Chapter 5.

Chapter 2

Flexible Models for Assessing Optimal Intervention Timing in patients with NSTE-ACSs

2.1 Introduction

Invasive intervention in patients with non–ST-segment elevation acute coronary syndromes (NSTE-ACSs) includes both, assessment procedures such as cardiac catheterization and therapies like revascularization. Assessment procedures are implemented first and are useful to decide which therapy to follow subsequently. Early execution of these type of interventions, usually before 72 hours since patient attendance, is established as the recommended treatment strategy, instead of following a conservative plan of drugs administration.

Nevertheless, as noted by Navarese et al., 2013, the optimal timing of intervention in NSTE-ACSs patients remains a matter open to debate. Some of the causes of the conflicting results are the use of different data sources (randomized-controlled trials versus observational data registry) and alternative risk profiles of populations under study. But additional sources of inconclusive results may be related to methodological issues, especially when observational registry data is exploited.

In observational studies, classic estimation procedures of treatment effects are exposed to bias and inconsistency problems due to residual confounding (also known as treatment endogeneity in the econometric literature). Therefore, regression methods based on instrumental variables are natural alternatives to handle the endogeneity bias problem.

Over last decade, several observational studies addressed this issue, such as Montalescot et al., 2005, Tricoci et al., 2007 and Sorajja et al., 2010, but all of them neglected the residual confounding problem. A notable exception is Ryan et al., 2005, which uses day of hospital presentation (weekend vs. weekday) as an instrumental variable (IV) to study the impact of the timing of cardiac catheterization and revascularization therapy over in-hospital mortality and other outcomes. They find non significant benefits for the early catheterization, although an important risk reduction cannot be excluded.

Following an identification strategy similar to that in Ryan et al., 2005, we study the impact of time delay to catheterization on outcomes for non ST-segment elevation Myocardial Infraction (NSTEMI) patients, exploiting the fact that patients admitted on weekends are less likely to undergo earlier catheterization than patients admitted during workweek days. Therefore, we employ this exogenous source of variation in the treatment *time delay to catheterization* to identify its causal effect on outcomes via regression models based on instrumental variables.

In contrast with the traditional approach (i.e. the one usually followed by researchers in this specific applied literature), employed by Ryan et al., 2005, we introduce innovations in two directions. On the one hand, we maintain the original continuous variable *time delay to catheterization* (*TDC*) as the relevant treatment, instead of specifying it as a binary (dummy) variable indicating *early catheterization*. On the other hand, our causal inference procedure relies on a flexible specification of the Triangular Simultaneous Equations Model, recently proposed by Marra and Radice, 2011.

The first innovation allows us to estimate a nonlinear effect of the continuous treatment variable *TDC* using a single two-dimensional system of triangular equations. To identify nonlinear effects, the traditional approach dichotomizes the continuous treatment *TDC*, specifying a set of binary treatment variables to indicate different levels of *early catheterization* (less than 12 hours, less than 24 hours, and so on) versus *late catheterization* (more/equal than 12 hours, more/equal than 24 hours, etc.), as is the case in Ryan et al., 2005. Such a representation of the treatment variable requires the estimation of a set of two-dimensional simultaneous equations systems (one for each binary treatment) when inference is based on instrumental variables. Furthermore, and more importantly, that kind of dichotomization over a continuous treatment variable can lead to an overestimation of the treatment effect, as is pointed at Baiocchi, Cheng, and Small, 2014 and showed by Angrist and Imbens, 1995.

The second innovation enables the estimation of smooth non-linear functions for both the treatment effect of *TDC* and the effects of continuous control variables. Additionally, it allows us to construct valid point-wise confidence intervals for the estimated smooth functions.

The remainder of the chapter is organized as follows. Section 2.2 describes the relevant sample of patients, defines the variables used and presents the general identification strategy. Section 2.3 specifies the alternative models to be estimated, which are classified into two main groups, those with an identity link function and those possessing a Probit link function. In Section 2.4 data description is presented and identification assumptions are assessed. Empirical results emerging from the estimation of all the considered models are exhibited at Section 2.5. Finally, Section 2.6 concludes.

2.2 Problem definition and identification strategy

The main objective of the following analysis consists of estimating the effect of time delay to catheterization (*TDC* hereafter) on outcomes, related to mortality and myocardial infraction, for NSTE-ACSs patients. Therefore, the outcome variable is defined by a binary variable called *Event*, indicating the presence of any of both situations: a) all-cause mortality from intervention to 12 months and b) acute myocardial infraction from intervention to 12 months.

The available sample includes NSTE-ACSs patients having undergone cardiac catheterization with a delay between 0 and nearly 1000 hours. This allowed us to measure the treatment delay in continuous time and to define the treatment variable *TDC*, measured in hours.

One of the primary obstacles to overcome involves the endogenous nature of *TDC*. This is so because the decision to catheterize is made based on patients characteristics which can be fully perceived by medical staff but are only partially observed by the researcher (due to its partial registration on data sources). Therefore, risk factors unobserved by the researcher determine both time to catheterization and the probability of outcome occurrence, causing the residual confounding problem.

If patient baseline characteristics and usual in-hospital treatments did not differ on the basis of weekday versus weekend presentation at hospital, weekend status could be used as a valid IV for assessing the effect of timing of cardiac catheterization on outcomes.

Weekend patients include those who presented to the admitting hospital between 5 pm Friday and 3 pm Sunday. All other patients were considered weekday patients. Then, the instrumental variables were defined as a set of three mutually exclusive dummy variables, indicating admissions on Friday, Saturday and Sunday. These definitions were chosen so as to maximize the number of weekend patients presenting more than 18 hours from presentation to Monday at 9 pm, at which time we expected catheterization laboratory facilities would be fully operational. This instruments definition differs from the one used by Ryan et al., 2005 and others applications, which only specify a unique binary variable indicating weekday patients, without distinction of specific weekend day. Nevertheless, using only one instrument does not significantly changes the main conclusions of the present analysis.

To focus on the effects of an *early* catheterization, we restrict the full sample to include patients who underwent catheterization within 60 hours from admission. Results remain the same if we restrict the maximum delay to 48 hours, as is the case in Tricoci et al., 2007, and times between 48 and 60 hours. One technical reason justifying the use of a bounded sample relates to the IVs requirements. Specifically, to better fulfill the instrumental variables condition related to having *enough partial correlation* with the endogenous variable *TDC*. In that sense, it is a logical and testable fact that partial correlation between IVs defined earlier and the treatment *TDC* decreases when the maximum time bound is increased.

Finally, we include two continuous control variables, *Age* (containing the patient age in years) and *Gr* (GRACE, Global Registry of Acute Coronary Events risk score) measured at hospital admission and the binary control *Fem* indicating female patients. Table 2.1 summarizes the relevant variables to be used.

TABLE 2.1: *Relevant variables included in the analysis*

Variable	Description
Event	The outcome binary variable. *Event*=1 indicates: a) all-cause mortality from intervention to 12 months or b) acute miocardial infraction from intervention to 12 months.
TDC	The continuous treatment variable *Time Delay to Catheterism*, measured in hours.
Age	Continuous control variable measuring Patient age in years.
Gr	Continuous control containing the Global Registry of Acute Coronary Events risk score (GRACE) measured at hospital admission.
Fem	Binary variable indicating female patients.
Fr	Binary instrumental variable indicating hospital admission at Friday.
Sa	Binary instrumental variable indicating hospital admission at Saturday.
Su	Binary instrumental variable indicating hospital admission at Sunday.

2.3 General specification and alternative models

According to the triangular simultaneous equations framework (Newey, Powell, and Vella, 1999), the general model of study can by represented by the following system,

$$Event = H_2(TDC, Age, Gr, Fem) + \varepsilon \tag{2.1}$$

$$TDC = H_1(Age, Gr, Fem, Fr, Sa, Su) + u \tag{2.2}$$

where (2.1) is the structural equation of main interest, linking the outcome *Event* with the treatment *TDC* through a nonparametric function $H_2(.)$, and adding *Age*, *Gr* and *Fem* as control variables. The reduced form equation (2.2) explains the endogenous treatment *TDC*, being exogenously affected by the binary instruments *Fr*, *Sa* and *Su* (indicating the three possible days for patients arriving on weekends) and depending potentially on *Age*, *Gr* and *Fem*. All regressors affect *TDC* through general functional form $H_1(.)$.

The foregoing is the full nonparametric version of the system, with additive separability of random errors (u and ε, which are intended as deviations from the mean functions H_2 and H_1) and identification assumptions given by

$$E(\varepsilon|u, Age, Gr, Fem, Fr, Sa, Su) = E(\varepsilon|u) \tag{2.3}$$

$$E(u|Age, Gr, Fem, Fr, Sa, Su) = 0. \tag{2.4}$$

Restriction (2.3) implies that unobserved risk factors ε are mean dependent of unobserved factors u affecting *TDC* and that such a relationship is unaffected by control variables and instruments (i.e. the whole set of exogenous independent variables in the system). On the other hand, condition (2.4) states that unobserved factors affecting *TDC* are mean independent of instruments and control variables.

Based on (2.1)-(2.4) and taking conditional expectation $E(Event|TDC, Age, Gr, Fem, Fr, Sa, Su)$ over (2.1), it is possible to specify the structural mean regression function as

$$\begin{aligned} Event &= E(Event|TDC, Age, Gr, Fem, Fr, Sa, Su) + \varepsilon \\ &= E(Event|TDC, Age, Gr, Fem, u) + \varepsilon \\ &= H_2(TDC, Age, Gr, Fem) + f_u(u) + \varepsilon \end{aligned} \tag{2.5}$$

where $f_u(u) = E(\varepsilon|u)$ represents the so called *control function* or *control variable* and $\varepsilon = \varepsilon - E(\varepsilon|u)$. Further, the reduced form regression function for equation (2.2) is

$$\begin{aligned} TDC &= E(TDC|Age, Gr, Fem, Fr, Sa, Su) + u \\ &= H_1(Age, Gr, Fem, Fr, Sa, Su) + u. \end{aligned} \tag{2.6}$$

Because *Event* is a dummy variable and its expected value represents the probability of occurrence of the event, we can redefine (2.6) to make explicit that we are estimating the probability model (2.7),

$$\begin{aligned} Event &= P(Event = 1|TDC, Age, Gr, Fem, u) + \varepsilon \\ &= H_2(TDC, Age, Gr, Fem) + f_u(u) + \varepsilon. \end{aligned} \tag{2.7}$$

Depending on how functional forms for H_1, H_2 and f_u are specified in (2.6) and (2.7), a range of alternative models arise. Here we focus on four models which can be classified in two groups. The first group ignores the binary condition of the outcome variable *Event*, and the second one takes it into account.

2.3.1 Linear and Additive Models

The most basic and frequently used specification is the parametric linear regression model, which raises the following two linear regression equations (2.8-2.9)

$$\begin{aligned} Event &= P(Event = 1|TDC, Age, Gr, Fem, u) + \varepsilon \\ &= \beta_0 + \beta_1 TDC + \beta_2 Age + \beta_3 Gr + \beta_4 Fem + \beta_5 u + \varepsilon \end{aligned} \tag{2.8}$$

$$\begin{aligned} TDC &= E(TDC|Age, Gr, Fem, Fr, Sa, Su) + u \\ &= \alpha_0 + \alpha_1 Age + \alpha_2 Gr + \alpha_3 Fem + \alpha_4 Fr + \alpha_5 Sa + \alpha_6 Su + u \end{aligned} \tag{2.9}$$

This model is mostly used in applications, despite it ignores the limited or bounded nature of the probability operator $P(Event = 1|...)$, for at least two reasons pointed out in Baiocchi, Cheng, and Small, 2014. First, because it often provides a good approximation to the average treatment effect through parameter β_1. Second, its specification does not require making parametric assumptions about the link function necessary to bound the behavior of operator $P(Event = 1|...)$, as is the case when popular non-linear probability models like Logit and Probit are used.

Its main practical disadvantage lies in the linear functional form imposed to the effect of the continuous treatment *TDC*. In this setting, one hour of increase in *TDC* affects the probability (or risk) of *Event* occurrence by the same amount (β_1) irrespective of the initial level of *TDC*. In fact, it can be argued that *TDC* could have a (positive) decreasing effect. In that case using the linear model would potentially lead to misguided generalizations about the size and significance of the treatment effect.

A straightforward way to relax the linearity assumption consists in specifying an additive regression model (Hastie and Tibshirani, 1986). In the present context, the additive representation can be stated as

$$\begin{aligned} Event &= P(Event = 1|TDC, Age, Gr, Fem, u) + \varepsilon \\ &= \beta_0 + f_{TDC}(TDC) + f_{Age}(Age) + f_{Gr}(Gr) + \beta_1 Fem + f_u(u) + \varepsilon \end{aligned} \tag{2.10}$$

$$\begin{aligned} TDC &= E(TDC|Age, Gr, Fem, Fr, Sa, Su) + u \\ &= \alpha_0 + g_{Age}(Age) + g_{Gr}(Gr) + \alpha_1 Fem + \alpha_2 Fr + \alpha_3 Sa + \alpha_4 Su + u \end{aligned} \tag{2.11}$$

where $f_x(x)$ and $g_x(x)$ are flexible smooth functions of regressor x. Expressions (2.10) and (2.11) constitutes a Nonparametric Triangular Simultaneous Equations Model with an additive structure (Newey, Powell, and Vella, 1999).

A common estimation issue in regression equations like (2.8) and (2.10) is the existence of heteroscedastic errors, caused by the variance structure of the binary dependent variable *Event*. Such a problem requires using some variance-covariance correction method, which is easily available for the linear case but puts a more challenging obstacle for inference in the additive model based triangular equations context.

The later type of heteroscedasticity can be avoided using a model which recognizes the binary nature of the dependent variable *Event*. Models of that kind are introduced in next subsection.

2.3.2 Generalized Linear Model and Generalized Additive Model

Another modeling alternatives arise when the bounded nature of $P(Event = 1|...)$ operator is taken into account. In that situation the concept of *link function* plays a relevant role, giving rise to the so called Generalized Linear Model (GLM), (Nelder and Wedderburn, 1972), and Generalized Additive Model (GAM), (Hastie and Tibshirani, 1986).

The features that distinguish GLM from GAM are the same which separates the Linear Model from the Additive Model (i.e. they point to distinction between *linear* versus *flexible* functional forms of the regressors effects).

Establishing the link function as $\iota(\cdot)$ and the response function as $l_0 = \iota^{-1}$ the GLM specification for the present application is given by (2.12) and (2.13),

$$\begin{aligned} Event &= P(Event = 1|TDC, Age, Gr, Fem, u) + \varepsilon \\ &= l_0(\beta_0 + \beta_1 TDC + \beta_2 Age + \beta_3 Gr + \beta_4 Fem + \beta_5 u) + \varepsilon \end{aligned} \tag{2.12}$$

$$\begin{aligned} TDC &= E(TDC|Age, Gr, Fem, Fr, Sa, Su) + u \\ &= \alpha_0 + \alpha_1 Age + \alpha_2 Gr + \alpha_3 Fem + \alpha_4 Fr + \alpha_5 Sa + \alpha_6 Su + u \end{aligned} \tag{2.13}$$

where ε represents the *response error*, i.e the difference between the dependent variable *Event* and its conditional expectation. When the response function $l_0(.)$ is specified as the Cumulative Density Function (c.d.f.) from a Standard Normal distribution, the regression probability model called Probit emerges. Alternatively, if a c.d.f. from a Standard Logistic distribution is used, the so called Logit Model is defined.

Logit is the most widely applied model in medicine and bio-statistics, partly because of its simple computation and its relation with the Odds Ratio measure of relative risk, very popular in discrete-treatment evaluation analyses.

We opted to apply the Probit link function because both, it is the most studied and applied model in the econometric literature (in the case of parametric probability models with endogenous regressors) and because computing odds ratios is not our goal due to the continuous nature of our treatment variable. Although this nonlinear parametric probability model does not constitutes the main interest for our analysis, it is an adequate benchmark to make comparisons. In any case, using Probit or Logit makes no difference on final conclusions.

On the other hand, the GAM specification including flexible smooth functions for each covariate is given by system (2.14)-(2.15),

$$\begin{aligned} Event &= P(Event = 1|TDC, Age, Gr, Fem, u) + \varepsilon \\ &= l_0(\beta_0 + f_{TDC}(TDC) + f_{Age}(Age) + f_{Gr}(Gr) + \beta_1 Fem + f_u(u)) + \varepsilon \end{aligned} \tag{2.14}$$

$$\begin{aligned} TDC &= E(TDC|Age, Gr, Fem, Fr, Sa, Su) + u \\ &= \alpha_0 + g_{Age}(Age) + g_{Gr}(Gr) + \alpha_1 Fem + \alpha_2 Fr + \alpha_3 Sa + \alpha_4 Su + u, \end{aligned} \tag{2.15}$$

where, as before, $l_0(.)$ represents the Probit response function of the binomial family.

Finally, it can be noted that reduced form equations 2.9 and 2.13 are identical, as are the equations 2.11 and 2.15.

2.4 Data description and assessment of identification assumptions

2.4.1 Sample and data description

Our data base includes patients admitted consecutively between November 2003 and January 2011 to the Cardiology Department from Clinic Hospital of Santiago, with an Acute Coronary Syndrome (ACS) diagnosis. The demographic and clinical data were collected prospectively and digitally recorded.

Patients were diagnosed with ACS if they showed new onset symptoms consistent with cardiac ischemia, cardiac biomarkers with values above the higher normal threshold and in any of the following events: electrocardiogram variation consistent with ACS, in-hospital stress testing suggesting ischemia, or registered history of coronary vessel disease.

Patients were classified as having STEMI (ST-segment elevation Myocardial Infraction) or NSTEMI (non ST-segment elevation Myocardial Infraction) or Unstable Angina. As was mentioned in the introduction only NSTEMI patients, catheterized before 60 hours since hospital admission, were selected for the present study. Patients whose ACS was precipitated in the context of surgery, sepsis, trauma, or cocaine consumption were excluded as well as patients presenting missing data for some control variable. As a result, 1101 patients constituted the final data base.

Table 2.2 summarizes baseline characteristics and instruments, presenting descriptive statistics for variables defined in Table 2.1, for patients in the final data set.

TABLE 2.2: *Relevant variables included in the analysis*

Variable	Mean	Std. Dev.	1st quar.	3rd quar.	Min	Max
Event	0.1453	0.3526	0	0	0	1
TDC	28.74	14.51	18	41	0	60
Age	64.65	12.11	55.75	74	33.7	92
Gr	124.6	39.15	98	148	33	290
Fem	0.2470	0.4315	0	0	0	1
Fr	0.0073	0.0850	0	0	0	1
Sa	0.0972	0.2963	0	0	0	1
Su	0.0872	0.2822	0	0	0	1

The event mortality/myocardial infraction affects 14.53% of patients in the sample, representing 160 cases. This is an usual rate for that kind of event. The mean value for treatment variable *TDC* rounds about 29 hours.

The control variables *Age* and *Gr* show mean values of around 65 years and 125, respectively. Female patients represent approximately 25% of sample patients, which is an usual rate.

Patients admissions rates are similar on Saturdays (8.7%) and Sundays (9.7%), and are relatively low on Fridays (0.7%) because only last 7 hours of that day are considered as weekend time. Final results and conclusions remains approximately the same when the dummy for Friday admission is excluded from the set of instruments or if a unique instrument (indicating weekend arriving at any day) is used.

2.4.2 Assessing identification assumptions

The asymptotic *Consistency* property for IV-based estimators depends on the satisfaction of key identifications assumptions, provided by equations (2.1)-(2.4), which are linked to instruments' validity. They can be summarized in the present application as follows:

A1) The set of instruments, *Fr*, *Sa* and *Su*, present partial (positive) correlation with endogenous treatment *TDC*, feature which is reflected in the reduced form equation 2.2, where the instruments are included as covariates.

A2) The instruments, *Fr*, *Sa* and *Su*, are partially uncorrelated with u (the unobserved factors affecting *TDC*) and partially uncorrelated with ε (the unobserved factors affecting *Event*), once conditioning on u. Both conditions are formally stated by equations (2.4) and (2.3), respectively.

A3) The instruments, *Fr*, *Sa* and *Su*, are not relevant explanatory variables in the structural equation (2.1), or in its alternative version (2.5). Then, the unique channel through which instruments can affect outcome is the treatment *TDC*.

It is important to remember that *TDC* was bounded to a maximum of 60 hours and the final sample includes only catheterized patients, which is a strategy designed to focus on early catheterization effect over outcomes, similar to that followed in Tricoci et al., 2007.

It must be noted that changing the 60 hours limit for *TDC*, affects its partial correlation with the instruments (*Fr*, *Sa* and *Su* dummy variables). In fact, a higher bound for *TDC* implies a lower correlation, because the instruments become less relevant to affect (i.e. to increase) the new values (i.e. values larger than 60) of *TDC*.

Therefore, defining the maximum threshold for *TDC* affects assumption A1. Such assumption requires significant partial correlation between the endogenous treatment and the instruments.

To assess the fulfillment of condition A1, Table 2.3 reports estimation results obtained by regressing *TDC* over the three binary instruments for weekend patients and the remaining control variables, using a linear model specification (i.e. fitting the reduced form equation 2.9, which is the same that 2.13).

TABLE 2.3: *Linear regression of TDC on controls and instruments*

Covariate	**Coef.**	**Std. err.**	**t-statistic**	**P-value**
Age	0.3097	0.0447	6.92	0.0000
Gr	-0.0987	0.0137	7.22	0.0000
Fem	0.6957	0.9464	0.735	0.4624
Fr	18.47	4.7272	3.907	0.0000
Sa	17.54	1.3649	12.85	0.0000
Su	4.35	1.4258	3.051	0.0023
Interc.	18.64	2.1998	8.472	0.0000
Adjusted $R^2 = 0.164$; Sample Size: 1101				
Instruments exclusion restrictions test: F=59.9; p-value=0.000				

As can be seen in second and third columns of Table 2.3, the instruments *Fr*, *Sa* and *Su* are both quantitatively relevant and statistically significant. Moreover, F statistic for testing instruments exclusion restrictions is several times larger than 12 and 16, which are the recommended minimum F values to exclude a weak instruments scenario when using 2 and 5 instruments, respectively (see Stock, Wright, and Yogo, 2002 for a survey on weak instruments). In fact, the set of instruments explains most of the *TDC*'s variability given that adjusting the model without them produces an adjusted R^2 of about 0.03 (notably smaller than 0.164, the value reported in Table 2.3).

As was mentioned earlier, using of only one instrument indicating weekend arrival does not produces substantial changes in results. We opted to use three disjoint IVs because it allows us to induce more exogenous variation in the treatment *TDC*.[1] In fact, regressing *TDC* on a single aggregated instrument, and other controls, produces a R^2 equal to 0.129 which is smaller than reported at Table 2.3 (0.164). This larger exogenous variation may be important to assess a possible nonlinear effect of *TDC*. In fact, in our estimations we find a slightly greater nonlinearity when the three IVs are used.

Replication of the analysis reported in Table 2.3 but using an additive model framework (i.e adjusting regression function 2.11 or 2.15), that incorporates smooth effects for continuous controls *Age* and *Gr*, does not provide additional insights (adjusted R^2 and F statistic slightly increase to 0.165 and 59.2, respectively).

[1] It can be noted that the estimated coefficient of instrument *Su* (of about 4.3) is clearly smaller than corresponding coefficients for *Sa* and *Fr* (17.5 and 18.5, respectively).

Assumption A2 basically says that instruments are not related to unmeasured factors which simultaneously affect *TDC* (through u) and *Event* (by means of ε). Such an assumption is not directly testable.

An indirect way of assessing A2 consists in testing the existence of a relationship between instruments and *measured* risk factors (i.e. covariates *Age*, *Gr* and *Female*). If there is no relation between instruments and measured factors, then it can be argued that instruments are randomly assigned or generated.

Table 2.4 reports difference of means tests, performed over each measured factor, comparing *weekend patients* versus *weekday patients*.

TABLE 2.4: *Covariates means: weekend vs. weekday patients*

Covariate	**Weekend patients**	**Weekday patients**	**P-value**
*Age**	64.5	64.7	0.8386
*Gr***	130.4	123.3	0.0266
*Fem****	0.2275	0.2517	0.4637

* Two Sample t-test (equal variances)

** Welch-Aspin Two Sample t-test (unequal variances)

*** Two Sample test of proportions

The GRACE score is the only measured factor which presents a statistical significant (positive) association with arrival on weekend status. But it can be argued that the GRACE level difference ($130.4 - 123.3 = 7.1$) does not represent a significant difference from a medical point of view. If the same t-test of Table 2.4 is performed for *Gr* but using the whole sample of 2635 catheterized patients (i.e. without imposing a maximum bound for *TDC*), a smaller means difference is obtained ($129.3 - 124.6 = 4.7$) with a p-value of about 0.0738, suggesting a weaker association.

It should be noted that all previous difference of means tests were performed employing a unique instrumental variable indicating weekend patients versus weekday patients, instead of using the original set of instruments (binary variables *Fr*, *Sa* and *Su*). In other words, the instruments were aggregated into a single one. This is not a limitation because all IVs arise from a common type of data generating process (i.e hospital admission in some specific day of the week). Moreover, conclusions in Table 2.4 remain the same when original instruments *Fr*, *Sa* and *Su* are used.

An additional test is possible if we recognize that the partial correlation between instruments and treatment *TDC* can be broken, in a sub-sample of patients, when *TDC* values are above a large enough minimum bound. In such case, *exclusion restriction* assumption (A3) can be tested jointly with the assumption that *instruments are not related with unmeasured factors in* ε, by testing whether instruments are associated with outcome *Event* in the structural equation (see Baiocchi, Cheng, and Small, 2014 subsection 6.1).

In concrete, setting *TDC*'s minimum at 100 hours and its maximum at some large enough value, the relationship between IVs and *TDC* become statistically not significant. For example, Table 2.5 reports results for the same analysis presented in Table 2.3 but using a sample of 981 patients with *TDC* ranging between 100 and 360 hours.

TABLE 2.5: *Regressing TDC on IVs when TDC's range is* $[100, 360]$

Covariate	Coef.	Std. err.	t-statistic	P-value
Age	-0.2253	0.218	-1.033	0.3018
Gr	0.2076	0.060	3.429	0.0006
Fem	-2.422	4.00	0.605	0.5455
Fr	-8.6333	8.609	-1.003	0.3162
Sa	-7.926	6.679	-1.187	0.2356
Su	-0.9671	9.095	-0.106	0.9153
Interc.	161.40	11.58	13.938	0.0000
Adjusted $R^2 = 0.0108$; Sample Size: 981 Instruments exclusion restrictions test: F=0.753; p-value=0.5207				

As indicated by the F-statistic value reported in the last row of Table 2.5 (F=0.75 and p-value=0.52), the partial correlation between the IVs and the treatment in this new sample has disappeared.

Then we need to test whether the instruments are significant regressors in the structural equation, using the new sample of patients (Table 2.6). Results in Table 2.6 represent evidence against the hypothesis that IVs are relevant regressors in addition to the regressors in the structural equation (2.12), showing a *Chi*^{2}-statistic p-value of 0.5362 for instruments joint exclusion restriction test.

The same conclusion is obtained if we specify the structural equation as (2.8) and (2.14), and/or select different samples of patients (i.e. setting alternative minimum and maximum values for *TDC* so that it is not related with IVs). This represents additional empirical evidence supporting assumptions A2 and A3. However, this evidence and the procedures applied to obtain it can be criticized because it is usual that patients included in the new sample (who present a larger delay to catheterization) possess a lower risk of event occurrence than original patients. In such a case, can be argued that the new sample does not represents to the same population in terms of risk factors, creating a potential *selection bias* problem.

Finally, note that it is not theoretically justified to perform tests for over-identifying restrictions, as the Sargan Test. This is because instruments share the same type of data generating

TABLE 2.6: *Probit-type structural equation (2.12) including IVs as covariates*

Covariate	Coef.	Std. err.	z-statistic	P-value
TDC	-0.00102	0.00091	-1.112	0.2662
Age	0.00615	0.00631	0.975	0.3295
Gr	0.00883	0.00160	5.518	0.0000
Fem	-0.30527	0.11943	-2.556	0.0106
Fr	0.05722	0.24133	0.237	0.8126
Sa	0.15114	0.17816	0.848	0.3962
Su	-0.30714	0.27784	-1.105	0.2690
Interc.	-2.38921	0.38984	-6.129	0.0000

Adjusted $R^2 = 0.0774$; Sample Size: 981
Instruments exclusion restrictions test:
Deviance=2.1786; $Pr(> Chi2)$=0.5362

process, i.e arriving to hospital on a particular day of the week. Therefore, testing the *validity* of one instrument, supposing at the same time that the remaining IVs are *valid*, represents a logical contradiction.

2.5 Empirical results

To estimate the *TDC* effect on outcome *Event*, taking account of a potential endogeneity bias problem, can be approached using one of the triangular simultaneous equation models defined in Section 2.3. Those models differ in both, functional form of continuous covariates effects and the link function connecting binary outcome variable *Event* with regressors. Both features are associated with potential existence of non-linear covariates effects, so identification of a nonlinear *TDC* effect is of particular interest.

2.5.1 Estimation with Linear and Additive Models

The standard regression approach applied in practice is the Linear Model based system (2.8)-(2.9), which is usually estimated by Two-Stage Least Squares (2SLS) method. In this case, first stage involves estimation of reduced form function (2.9) leading to its adjusted version (2.16),

$$\begin{aligned} TDC_i = &\hat{\alpha}_0 + \hat{\alpha}_1 Age_i + \hat{\alpha}_2 Gr_i + \hat{\alpha}_3 Fem_i + \hat{\alpha}_4 Fr_i \\ &+ \hat{\alpha}_5 Sa_i + \hat{\alpha}_6 Su_i + \hat{u}_i,\ i = 1, ..., n; \end{aligned} \tag{2.16}$$

were subscript i indicates ith patient in sample of size n. The estimated coefficients in (2.16) was already reported in Table 2.3.

In common practice, the second step consists in estimating the structural equation that follows

$$Event_i = \hat{\beta}_0 + \hat{\beta}_1 T\hat{D}C_i + \hat{\beta}_2 Age_i + \hat{\beta}_3 Gr_i + \hat{\beta}_4 Fem_i + \varepsilon_i \tag{2.17}$$

were $T\hat{D}C$ is the vector of fitted values from the first stage estimated function (2.16), which is included as a regressor instead of the original treatment regressor *TDC*.

When this linear model setting is used, the same estimated coefficients produced in (2.17) can be obtained by fitting the structural equation as was defined in (2.8), leading to its adjusted version (2.18),

$$Event_i = \hat{\beta}_0 + \hat{\beta}_1 TDC_i + \hat{\beta}_2 Age_i + \hat{\beta}_3 Gr_i + \hat{\beta}_4 Fem_i + \beta_5 \hat{u}_i + \hat{\varepsilon}_i, \tag{2.18}$$

were $\hat{u}_i$ is the first stage residual of *i*th patient. These two equations, (2.18) and (2.16), constitutes the Control Function Approach (CFA) which was described in Section 2.3. When we moved from the parametric linear setting to the flexible semiparametric models such as GAMs, the CFA becomes the workhorse estimation procedure to deal with triangular simultaneous equations models in the frequentist framework, as we will see later.

Table (2.7) presents the estimated values of second stage coefficients in (2.17) or (2.18), with corresponding standard errors (non-robust and robust to heteroscedasticity) and 95% confidence intervals based on asymptotic normality.

TABLE 2.7: *Linear structural equation adjusted by 2SLS*

Covariate	**Coef.**	**Std. err.**	**Robust** **Std. err.***	**P-value** **P$>|z|$**	**95%** **Conf. Int.**
TDC	0.0048	0.0019	0.0022	0.0329	[0.0004, 0.0092]
Age	0.0009	0.0013	0.0013	0.4711	[-0.0033, 0.0015]
Gr	0.0029	0.0004	0.0004	0.0000	[0.0021, 0.0038]
Fem	-0.0879	0.0246	0.0223	0.0000	[-0.1315, -0.0442]
Interc.	-0.2799	0.0706	0.0745	0.0002	[-0.4719, -0.0880]

Adjusted $R^2 = 0.0453$; Sample Size: 1101
Breusch-Pagan/Cook-Weisberg heteroscedasticity test: $Pr(> Chi2)$=0.0000
*Hubert-White sandwich type, heteroscedasticicty robust, standard errors

First, results show a point estimation for *TDC* effect (β_1) equal to 0.0048, implying that one hour increment in time delay to catheterization increases 0.48% the probability of event occurrence (i.e. death or myocardial infraction). This is a medically significant effect which means, for example, that one day of delay to catheterization represents an 11,5% increase in event risk.

Even though inference must be based on heteroscedasticity-robust standard errors, because homoscedasticity hypothesis was rejected (see reported Breusch-Pagan test), we additionally reported the non-robust versions to show that conclusions remain roughly the same without standard errors correction. Coefficients of control variables GRACE (*Gr*) and female (*Fem*) are highly statistically significant and present the expected signs.

To observe the impact of the endogeneity correction performed by previous 2SLS procedure, Table (2.8) reports estimation results for the structural equation using Ordinary Least Squares (OLS) assuming the treatment TDC is exogenous.

TABLE 2.8: *Linear structural equation adjusted by OLS*

			Robust	**P-value**	**95%**
Covariate	**Coef.**	**Std. err.**	**Std. err.***	**P> $\|z\|$**	**Conf. Int.**
TDC	-0.0004	0.0007	0.0007	0.552	[-0.0017, 0.0009]
Age	0.0004	0.0011	0.0011	0.696	[-0.0018, 0.0027]
Gr	0.0025	0.0003	0.0004	0.0000	[0.0017, 0.0033]
Fem	-0.0851	0.0239	0.0218	0.0000	[-0.1278, -0.0424]
Interc.	-0.1672	0.05750	0.0564	0.003	[-0.2778, -0.0566]

Adjusted $R^2 = 0.09$; Sample Size: 1101
Breusch-Pagan/Cook-Weisberg heteroscedasticity test: $Pr(> Chi2)$=0.0000
*Hubert-White sandwich type, heteroscedasticicty robust, standard errors

As can be seen, the OLS estimation of the *TDC* effect β_1 is about -0.0004 and statistically insignificant. Comparing with 2SLS estimation ($\hat{\beta}_1 = 0.0048$) it is clear that the estimated difference between methods is medically relevant.

A formal assessment of the *TDC* exogeneity assumption is given by the Haussman test (Wald type test) which tests the significance of the coefficient of $\hat{u}_i$, β_5, in the structural equation (2.18). This test reveals evidence against the exogeneity assumption, producing a t-statistic equal to -2,615 (p-value = 0.009) for β_5 and suggesting 2SLS as the preferred method.

As explained earlier in subsection 2.3.1, a limitation of the previously estimated triangular system is related to the linear specification of the regressors' effects. For example, this assumption implies that a marginal increase of one hour in *TDC* produces the same marginal effect (β_1) across the whole range of time delay values (i.e. imply a constant marginal effect irrespective of the starting value for *TDC*). Imposing *effect constancy* may hide a nonlinear behavior relevant for medical practice.

One alternative that allows the existence of general non-linearities in regressors effects, consists in estimating an additive model-based triangular system given by equations (2.10) and (2.11). One recent estimation method for such flexible system, proposed by Marra and Radice, 2011 and called Two Stage Generalized Additive Model (2SGAM), employs a two steps procedure similar to the CFA, used earlier.

Now, in the first stage we need the estimated version of the reduced form regression (2.11), defined as

$$\begin{aligned} TDC_i =& \hat{\alpha}_0 + \hat{g}_{Age}(Age_i) + \hat{g}_{Gr}(Gr_i) + \hat{\alpha}_1 Fem_i \\ & + \hat{\alpha}_2 Fr_i + \hat{\alpha}_3 Sa_i + \hat{\alpha}_4 Su_i + \hat{u}_i,\ i = 1, \ldots, n. \end{aligned} \tag{2.19}$$

The second stage involves estimating structural equation (2.10), using first stage residuals $\hat{u}$ as a regressor instead of unobserved errors u, this can be expressed as

$$\begin{aligned} Event_i =& \hat{\beta}_0 + \hat{f}_{TDC}(TDC_i) + \hat{f}_{Age}(Age_i) + \hat{f}_{Gr}(Gr_i) \\ & + \hat{\beta}_1 Fem_i + \hat{f}_{\hat{u}}(\hat{u}_i) + \hat{\varepsilon}_i. \end{aligned} \tag{2.20}$$

The *naive* AM (Additive Model) estimation, without correcting for endogeneity, would consist in estimating (2.20) excluding $\hat{u}$ as regressor.

TABLE 2.9: *Estimation results for Aditive Model-based first stage (2.19)*

Covariate	(e.d.f.)/coef.	Std. err.	(F)/t- statistic	P-value
$\hat{g}_{Age}(Age)$	(1.001)	–	(29.7)	0.0000
$\hat{g}_{Gr}(Gr)$	(1.402)	–	(47.9)	0.0000
Fem	0.7009	0.9463	0.741	0.4591
Fr	18.50	4.7267	3.914	0.0000
Sa	17.54	1.3647	12.86	0.0000
Su	4.34	1.4257	3.045	0.0024
Intercept	26.35	0.5038	52.31	0.0000

Adjusted $R^2 = 0.165$; Sample Size: 1101
Instruments exclusion restrictions test: F=53.2; p-value=0.000

Tables 2.9 and 2.10 report estimation results for first (2.19) and second (2.20) stages respectively, showing estimated coefficients for binary regressors and empirical degrees of freedom (e.d.f) of estimated smooth effects for continuous covariates.

As mentioned in subsection 2.4.2, adding flexible smooth effects in the first stage does not provide additional explanatory power when compared to the linear model case. In fact, estimated smooth effects of *Age* and *Gr* are practically linear, as indicated by respective e.d.f. in Table 2.9.

On the other hand, estimation in second stage shows a clear nonlinear effect for treatment *TDC* and GRACE score *Gr*, presenting e.d.f. about 2.5 and 4.1 respectively (Table 2.10). We will henceforth call this two-step estimation 2SAM (Two Stage Additive Model), because both dependent variables are treated as continuous.

TABLE 2.10: *Structural equation (2.20) adjusted by 2SAM*

Covariate	(e.d.f.)/coef.	Robust Std. err.*	95% Bayesian Conf. Int.*
$\hat{f}_{TDC}(TDC)$	(2.457)	–	–
$\hat{f}_{Age}(Age)$	(1.370)	–	–
$\hat{f}_{Gr}(Gr)$	(4.132)	–	–
$\hat{g}_u(\hat{u})$	(1.000)	–	–
Fem	-0.0862	0.021	[-0.1258, -0.0442]
Intercept	0.1666	0.011	[0.1438, 0.1887]

Smoothing parameter estimation by REML
Sample Size: 1101; REML score = 353.13
*Std. errors and Bayesian C.I. derived from simulation
and corrected for heteroscedasticity trough weighting.

Figure 2.1 presents estimations for *TDC* smooth effect (left), for both 2SAM and AM (i.e. without endogeneity correction), and its first derivative or marginal effect (right) for the 2SAM case. Based on 95% bayesian confidence intervals, naive AM estimates seems to be non-significant while 2SAM case shows a (positive) marginal effect (*MgEf*) that is significant between 0 and 34 hours only. Thus, the significant *MgEf*'s values range approximately from 0.011 to 0.005, showing a relevant heterogeneity from a medical perspective.

Previous 2SAM results expose how misleading a careless usage of 2SLS could be, as it implies a significant *MgEf* of 0.0048 for the whole *TDC* range (i.e from 0 to 60 hours). Based on 2SAM we can compute a global *MgEf* (i.e. average first derivative) of about 0.0059 averaging over the whole *TDC* range. Moreover, averaging over the statistically significant *TDC*'s values

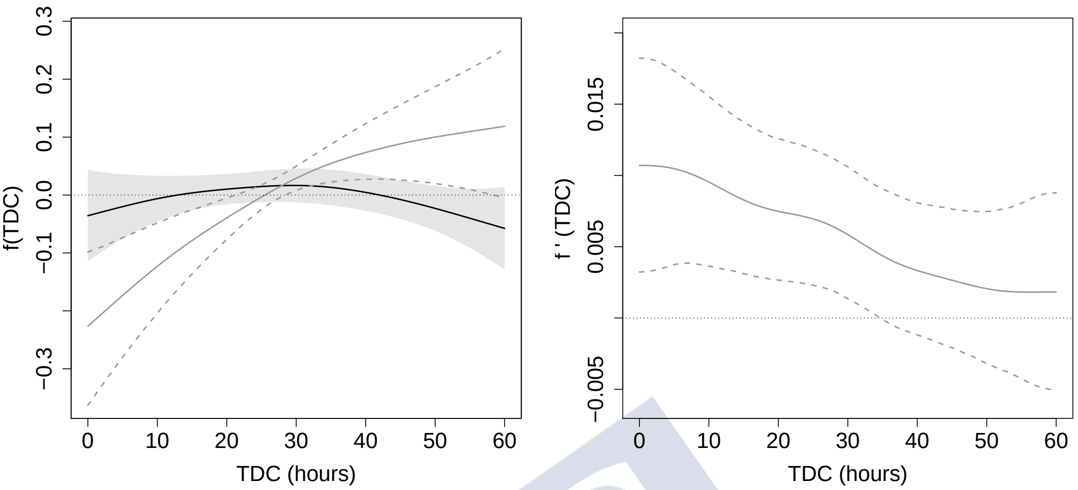

FIGURE 2.1: *Left: smooth TDC effects estimated by 2SAM (grey line) and AM (black line), with respective 95% bayesian C.I. (dashed line for 2SAM and shaded area for AM). Right: estimated marginal effect (first derivative) for TDC and corresponding 95% C.I. using 2SAM.*

(i.e averaging from 0 to 34 hours), we get a larger *MgEf* of about 0.0076. Both instances demonstrate a medically relevant *MgEf* sub-estimation from 2SLS approach.

Estimated smooth effects for *Age*, *Gr* and first stage residuals ($\hat{u}$) are plotted in Figure 2.2. The *Age* effect becomes completely non-significant when estimated by 2SAM and *Gr* possesses nearly the same nonlinear effect at 2SAM than at naive AM.

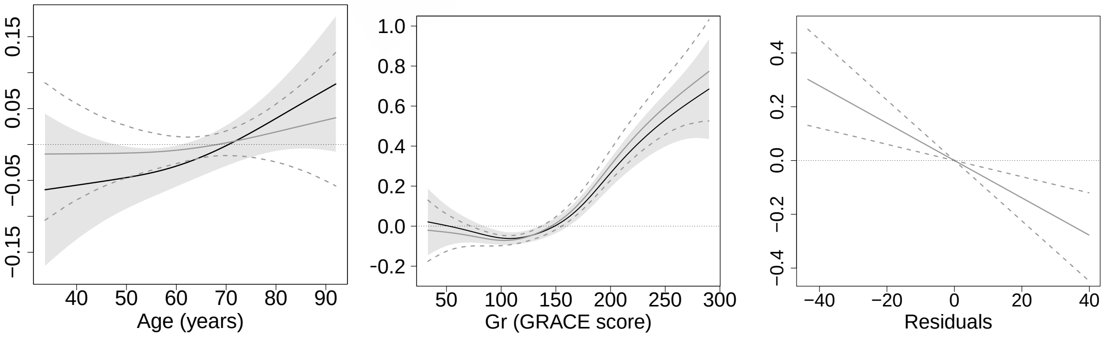

FIGURE 2.2: *Smooth effects estimations, using AM (black line) and 2SAM (grey line) for Age, Gr and Residuals at structural equation with additive modeling, with respective 95% bayesian C.I. (shaded area for AM and dashed lines for 2SAM).*

It is important to note that the first stage residuals seem to have a non-zero negative linear effect in the structural equation. In fact, using a Wald test statistic proposed by Wood, 2013 and used in Zanin, Radice, and Marra, 2014 it is possible to reject the null hypothesis of zero effect associated with first stage residuals (which is equivalent to rejecting the hypothesis of *TDC* exogeneity) with an F-statistic of 9.46 and corresponding p-value equal to 0.0022. Such a test is an extension of the procedure for testing the hypothesis of *TDC* exogeneity, usually called Hausman test of endogeneity (due to Hausman, 1978), used in the parametric linear case (Wooldridge, 2010).

More details about estimation and inference, such as smoothing parameter selection and heteroscedasticity corrections, and code for results replication are presented in the appendix chapter.

2.5.2 Estimation with GLM and GAM

The empirical analysis presented in the previous subsection does not recognize the binary nature of the dependent variable *Event* in the structural equation. A first step to introduce this binary condition consists in estimating a triangular system with the form in (2.13) - (2.12), i.e using a Generalized Linear Model to fit the structural equation (2.12).

In this context, estimation can be performed with the same two stage procedure, based on the Control Function Approach, described in previous subsection (2SGLM hereafter). Earlier works handling related issues, based on a Probit-type link function for the GLM model, are Heckman, 1978 and Amemiya, 1978. We applied the two stage estimation procedure proposed by Newey, 1987, which constitutes the standard parametric procedure when the structural equation is specified as a Probit model and the reduced form equation has a continuous dependent variable.

The reduced form equation is the same as in the linear model case; then, its estimated version, obtained in the first stage, is represented by (2.16). On the other hand, the second stage estimated structural equation is given by

$$Event_i = l_0(\hat{\beta}_0 + \hat{\beta}_1 TDC_i + \hat{\beta}_2 Age_i + \hat{\beta}_3 Gr_i + \hat{\beta}_4 Fem_i + \beta_5 \hat{u}_i) + \hat{\varepsilon}_i \tag{2.21}$$

where l_0 is the Probit response function and $\hat{\varepsilon}_i$ is the response residual for ith patient.

Coefficients in equation (2.21), estimated by 2SGLM as in Newey, 1987, are reported in Table 2.11. *TDC* has a 95% significant effect, showing an estimated coefficient equal to 0.0207 which cannot be interpreted beyond its sign. Again, *Gr* and *Fem* are significant controls while *Age* is not.

TABLE 2.11: *Structural equation adjusted by 2SGLM*

Covariate	Coef.	Std. err.	P-value P> $\|z\|$	95% Conf. Int.
TDC	0.0207	0.0086	0.0160	[0.0039, 0.0376]
Age	-0.0009	0.0061	0.8860	[-0.0127, 0.0110]
Gr	0.0115	0.0018	0.0000	[0.0080, 0.0150]
Fem	-0.4436	0.1293	0.0000	[-0.6972, -0.1901]
Interc.	-3.0279	0.3521	0.0000	[-3.7181, -2.3378]

Sample Size: 1101
Wald test of exogeneity: $Chi2(1)$ = 8.67 $Pr(> Chi2)$=0.0032

In this GLM setting, the marginal effect (*MgEf*) of the continuous treatment *TDC* depends on the others independent variables. Defining the estimated probability as follows

$$\begin{aligned}\hat{P}(Event_i = 1|TDC, Age, Gr, Fem, \hat{u}) =& l_0(\hat{\beta}_0 + \hat{\beta}_1 TDC_i + \hat{\beta}_2 Age_i \\ &+ \hat{\beta}_3 Gr_i + \hat{\beta}_4 Fem_i + \beta_5 \hat{u}_i),\end{aligned} \tag{2.22}$$

and taking first derivative of *TDC* at (2.22) we obtain the *TDS*'s *MgEf* (2.23),

$$\begin{aligned}\frac{\partial \hat{P}(Event_i = 1|TDC, ..., \hat{u})}{\partial TDC_i} =& l_0'(\hat{\beta}_0 + \hat{\beta}_1 TDC_i + \hat{\beta}_2 Age_i \\ &+ \hat{\beta}_3 Gr_i + \hat{\beta}_4 Fem_i + \beta_5 \hat{u}_i)\hat{\beta}_1,\end{aligned} \tag{2.23}$$

where l_0' is the first derivative of the Probit response function (i.e the standard normal density function), which is a non-negative nonlinear function of *TDC* and the others regressors

Averaging (2.23) through *i* gives an estimated average *MgEf* of about 0.045, a value similar to the constant effect given by 2SLS (0.048).

The Wald test of exogeneity, testing the significance of first stage residuals at the second stage, rejects the null hypothesis of exogeneity (Table 2.11 last row). In fact running naive Probit regression on structural equation, assuming *TDC* exogeneity, gives a non-significant (p-value of 0.614) estimated coefficient β_1 of 0.0018 (not reported).

Finally, the more flexible triangular system specification, given by (2.14)-(2.15), can be estimated by 2SGAM (Marra and Radice, 2011). This is the same procedure used in the previous subsection to estimate the system based on additive models, in fact the first stage is exactly the same (see equation 2.19 and Table 2.9). The novelty in this case is the estimated structural equation given by (2.24).

$$\begin{aligned} Event_i =& l_0(\hat{\beta}_0 + \hat{f}_{TDC}(TDC_i) + \hat{f}_{Age}(Age_i) + \hat{f}_{Gr}(Gr_i) \\ &+ \hat{\beta}_1 Fem_i + \hat{f}_{\hat{u}}(\hat{u}_i)) + \hat{\varepsilon}_i, \end{aligned} \tag{2.24}$$

where l_0 is again the Probit response function.

Table 2.12 and Figures 2.3 present estimation results for (2.24). Estimated smooth functions and coefficients show a behavior parallel to those estimated with the additive model (2.20), but now they are not directly interpretable in terms of outcome probability.

TABLE 2.12: *Structural equation (2.24) adjusted by 2SGAM*

Covariate	**(e.d.f.)/coef.**	**Std. err.**	**95% Bayesian Conf. Int.***
$\hat{f}_{TDC}(TDC)$	(2.690)	–	–
$\hat{f}_{Age}(Age)$	(1.320)	–	–
$\hat{f}_{Gr}(Gr)$	(3.479)	–	–
$\hat{g}_u(\hat{u})$	(1.000)	–	–
Fem	-0.4636	0.1343	[-0.7269, -0.2001]
Intercept	-1.0595	0.057	[-1.1712, -0.9464]

Smoothing parameter estimation by REML
Sample Size: 1101; REML score = 353.13
*Std. errors and Bayesian C.I. derived from simulation

A small discrepancy with respect to the 2SAM case can be seen for the *TDC* smooth effect (Figure 2.3), which presents a slightly negative first derivative for the 52-60 hours range (but statistically non significant). Again, the naive GAM estimation produces a non-significant effect (black line with shaded area for 95% C. I.), and the Wald test rejects the exogeneity hypothesis with F-statistic equal to 11.52 and p-value of 0.0007.

As in the case of 2SGLM, the estimated probability now is given by

$$\begin{aligned} \hat{P}(Event_i = 1|TDC, ..., \hat{u}) =& l_0(\hat{\beta}_0 + \hat{f}_{TDC}(TDC_i) + \hat{f}_{Age}(Age_i) \\ &+ \hat{f}_{Gr}(Gr_i) + \hat{\beta}_1 Fem_i + \hat{f}_{\hat{u}}(\hat{u}_i)), \end{aligned} \tag{2.25}$$

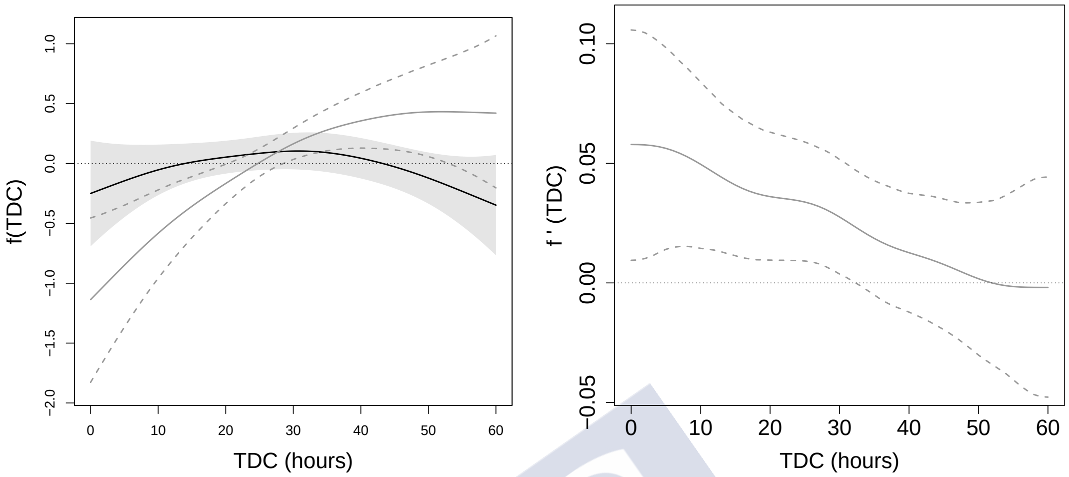

FIGURE 2.3: *Left: smooth TDC effects estimated by 2SGAM (grey line) and GAM (black line), with respective 95% bayesian C.I. (dashed line for 2SGAM and shaded area for GAM). Right: estimated first derivative of $\hat{f}_{TDC}(TDC)$ and corresponding 95% C.I. using 2SGAM.*

where the Probit type response function l_0 impose a kind of interaction between regressors, implying a nonlinear structure on *TDC*'s marginal effect, which can be expressed as

$$\begin{aligned}\frac{\partial \hat{P}(Event_i = 1|TDC,...,\hat{u})}{\partial TDC_i} =& l_0'(\hat{\beta}_0 + \hat{f}_{TDC}(TDC_i) + \hat{f}_{Age}(Age_i) \\ &+ \hat{f}_{Gr}(Gr_i) + \hat{\beta}_1 Fem_i + \hat{f}_{\hat{u}}(\hat{u}_i))\hat{f'}_{TDC}(TDC_i)\end{aligned} \tag{2.26}$$

where l_0' is the standard normal density function and $\hat{f'}_{TDC}(TDC_i)$ is the first derivative of *TDC* estimated smooth term (right graph in Figure 2.3).

Expression (2.26) can be used to evaluate *TDC*'s *MgEf* at different risk levels. For example, fixing all smooth effects $\hat{f}_x(x)$, except $\hat{f}_{TDC}(TDC)$, and linear term $\hat{\beta}_1 Fem_i$ at their median values we get *TDC*'s *MgEf* function evaluated at a kind of 'median risk' patient. Such a case is represented in Figure 2.4 (left) jointly with corresponding 'median risk' *MgEf* function for the 2SGLM case. Additionally, right graph in Figure 2.4 represents estimated probability (2.25) as a function of *TDC*, evaluated at the 'median risk' patient, for both 2SGAM and 2SGLM.

The main difference between 2SGAM and 2SGLM is that the later implies a 95% significant *MgEf* over the whole *TDC* range, as we concluded from Table 2.11, while 2SGAM produces a non-zero *MgEf* (based on a 95% confidence interval) over the smaller range located between 0

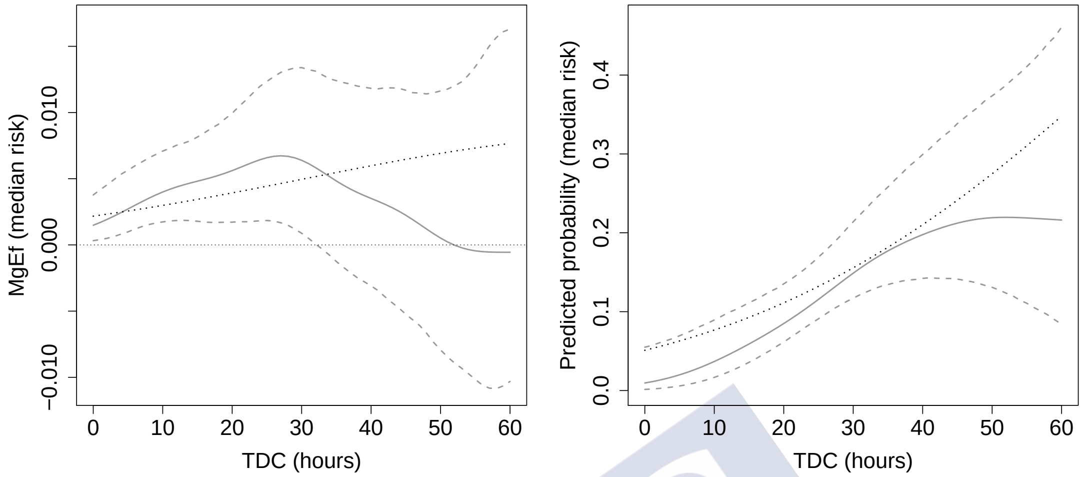

FIGURE 2.4: *Left: TDC's MgEf estimated by 2SGAM (grey line) with respective 95% bayesian C.I. (grey dashed line) and by 2SGLM (black dotted line). Right: predicted probability as a function of TDC estimated by 2SGAM (grey line) with corresponding 95% C.I. (dashed grey line) and by 2SGLM (dotted black line).*

and 32 hours. Over this smaller range, the *MgEf* takes values between 0.0067 and 0.0015. This is the same feature that explains the difference between the linear based 2SLS and the additive based 2SAM, as commented in previous subsection. Therefore, this feature can be attributed to the flexible structure of regressors effects in both AM and GAM cases, independently of the link function used at second stages (i.e. identity function or Probit).

The sample average of expression (2.26) gives a global *MgEf* of about 0.0055, which is larger than its analog from 2SGLM (0.0045) and similar to the one obtained using 2SAM (0.0059). Moreover, averaging (2.26) over *TDC* values from 0 to 32 (i.e. the grid were the effects are different from zero based on 95% C.I.) gives an average *MgEf* equal to 0.0076 (the same value obtained in the 2SAM case).

2.6 Discussion

In this chapter we have compared different modeling strategies, from parametric to semi-parametric regression models, applied to assessing optimal time of catheterization in NSTE-ACSs patients.

From a methodological point of view two main contributions to the existing literature were considered. In the first place, we used the treatment *TDC* ('Time Delay to Catheterism') as a continuous variable, instead of transforming it into dichotomous indicators of 'early intervention'. In the second place, we employed a recent flexible estimation procedure (Two Stage

Generalized Additive Model), based on the Triangular Simultaneous Equations Model, to account for both the presence of endogeneity bias (confounding) and the existence of nonlinear regressors effects.

Both combined innovations allowed us to estimate a flexible function for the treatment's marginal effect *TDC*. This function enables us to assess the relevance of the effect across the entire range of treatment values.

Another innovation was the construction of three binary instrumental variables, indicating the specific day of the weekend that patients arrived to the hospital, while the usual approach in the literature consists in using a single instrument indicating arrival on weekend. That allowed us to induce a larger exogenous variation to the endogenous treatment *TDC*, improving the fulfillment of identification assumptions by enhancing the instruments' strength.

Endogeneity bias was found to be a major concern, causing that the naive regression models completely fail to identify any significant treatment effect.

From a medical perspective, results support the existence of a nonlinear positive effect of *TDC* on patients survival and health status. Moreover, flexible modeling permits identification of a specific range for *TDC* values, from 0 to 30 hours approximately, in which treatment effect shows a nonzero (i.e. statistically significant) marginal effect which varies between 0.011 and a value slightly above 0. But the form in which this *MgEf* is related to *TDC* differs between 2SAM and 2SGAM estimation alternatives, as can be seen in Figure 2.5.

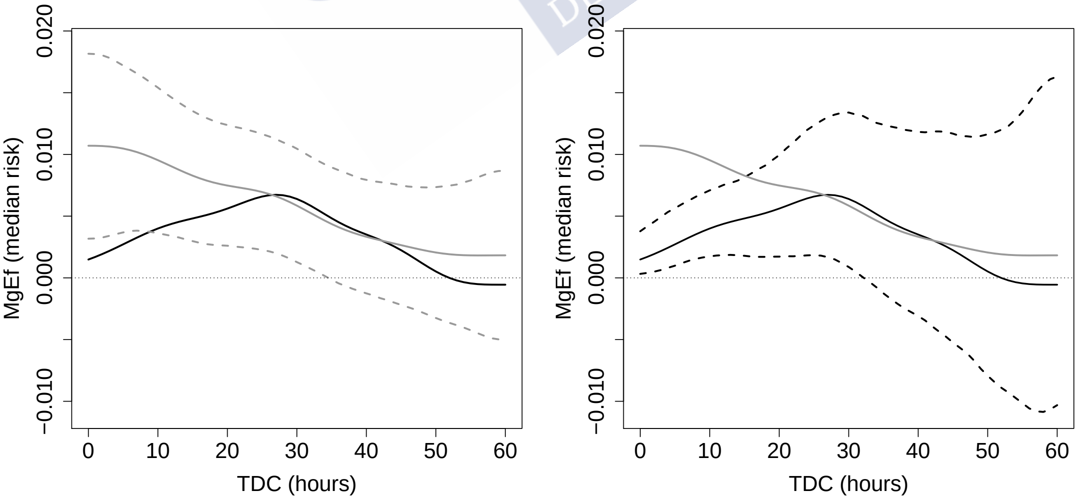

FIGURE 2.5: *Left: TDC's MgEf estimated by 2SAM (grey line) with respective 95% bayesian C.I. (grey dashed line) and by 2SGAM (black line). Right: TDC's MgEf estimated by 2SGAM (black line) with respective 95% bayesian C.I. (black dashed line) and by 2SAM (grey line).*

In the range from 0 to 28 hours, 2SAM estimation presents a positive decreasing *MgEf* while 2SGAM produces a positive increasing one. Such difference between marginal effects is due to the presence (absence) of a Probit link function, which imposes a specific functional form to predicted probabilities and their first derivative. When we consider the 95% confidence intervals in Figure 2.5, they are too wide to conclude that *MgEf* functions can be different in general, except inside the range from 0 to 15 hours.

Probit and Logit link functions (i.e cumulative density functions of Standard Normal and Logistic distributions), have first derivatives that vanish when extreme values of the link are reached. This feature supposes, in such extremes, that marginal changes in treatment (or another regressor) do not involve medically significant variations in cardiac complications. Such an assumption is not necessarily appropriate from a medical point of view.

Beyond the differences detected, both models bring empirical evidence supporting that early catheteriztion is a good decision within the first 30 hours since hospital admission, and the earlier, the better.

A possible extension for future research would consist in specifying the structural equation using a GAM with an unknown link function. This extension avoids the requirement of using a fixed parametric link function, which can be the source of biased results if it strongly differs from the true link. Estimation methods of such model, in the context of a single regression equation, was addressed by Roca-Pardiñas et al., 2004, Cadarso-Suárez et al., 2005, Horowitz and Mammen, 2011 and Tutz and Petry, 2013. No work was found studying this extension in the simultaneous equations framework.

Chapter 3

Identifying Class Size Effect on Schooling Achievement trough Flexible Triangular Equations Models

3.1 Introduction

Improving the quality of education at primary schools is an essential public policy goal in most developing countries. In these countries, especially in Latin American ones, a large and increasing amount of financial resources are destined to public schools maintenance. Such a large financial support represents an issue to policy debate and it is usually justified because strategic status of basic education. These issues led social researchers to study the impact of several factors on student achievement, mainly comparing effects due to *school* characteristics and *family* characteristics. The main goals of those works was to determine whether improving schools resources, quantitatively and/or qualitatively, can produce a relevant improvement on students' performance.

One of the most prominent resources that policymakers can alter is the *number of teachers* which generates a corresponding variation in *class size* (henceforth CS). The effect of class size reduction on student achievement was a highly studied and debated issue in the United States during the'80 and '90 decades, see for example Mishel et al., 2002 for a review of main findings. These studies found mixed results which imply a lack of unanimous evidence against or in favor of CS reduction; instead, CS affects achievement only within specific sub-populations of schools. More recently, the effects of class size reduction was assessed for European and others countries, many of them exploiting international surveys as the Programme for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS); see Woessmann, 2006, Hanushek and Woessmann, 2017 and Cordero, Cristóbal, and Santín, 2017 for recent reviews.

Class size *reduction* are usually supported by many social actors. Students' parents like small classes because they enable teachers to pay more attention to each student. Also, teachers can better handle group behavior, which is crucial for implementing learning methods and minimizing class disruptions. Teachers may prefer small classes because they imply fewer efforts in executing instruction process. And teacher unions presumably like class reduction since it requires additional teachers.

From a behavioral approach, there are recent theoretical efforts that try to explain the possible effects of class size. One instance is the *disruption model* of educational production developed in Lazear, 2001. This model implies both that *optimal* CS is larger for better-behaved pupils and that CS effect is bigger in groups of worst-behaved students. Such conclusion helps to explain actual difficulties to find conclusive empirical evidence through usual statistical techniques. For example, if students are allocated into classes of different size by schools administrators, in an attempt to reach optimal scholastic results, then random assignment of pupils into classes is broken. Therefore, a mayor problem to deal with when trying to measure CS effect on scholastic achievement is the endogenous status of CS.

One of the earlier empirical studies confronting class size endogeneity using instrumental variables methods, and linking this method with Fuzzy Regression Discontinuity Design, is Angrist and Lavy, 1999. Angrist and Lavy's pioneering work exploits Israeli's regulations, which impose a ceiling of 40 pupils per class, to obtain exogenous variation in CS.

Several subsequent works, dealing whit the class size effect on pupil performance, employed the identification strategy proposed by Angrist and Lavy, 1999 and reported mixed findings. For example, Bonesrønning, 2003 finds a significant CS effect in Norway secondary schools. Woessmann, 2005 finds zero effect in European countries using TIMMS survey. Urquiola, 2006 addresses the case of rural schools in Bolivia finding significant positive effects of CS reduction. In other study for Norway, Leuven, Oosterbeek, and Rønning, 2008 finds no evidence of a CS effect. Analyzing data from French junior high schools, Gary-Bobo and Mahjoub, 2013 found a significant but rather small effect of class size. Using TIMMS data for primary schools in Cyprus, Konstantopoulos and Shen, 2016 found both significant and non-significant CS effects, depending on the grade evaluated. Finally, analyzing data from TIMMS 2011 for 14 European countries, Li and Konstantopoulos, 2016 does not find systematic patterns of class size effects across countries, with the exception of Romania and the Slovak Republic.

Practical implementation of this specific identification strategy presents several complexities in terms of estimation and inference procedures. First, correct estimation of the relationship between the endogenous variable (CS) and the instrumental variable is often affected by the presence of outlier observations. Second, the presence of hierarchical data structures represents a challenge to the validity of standard inferential methods based on both parametric structures and asymptotic normality assumptions. In third place, relevant control variables affecting student achievement usually posses a non-linear effect.

To tackle such complexities, this chapter presents three procedural deviations to the standard instrumental variables method described in Angrist and Lavy, 1999 and used in the articles previously cited. As a first innovation we propose a flexible triangular equations model, recently proposed by Marra and Radice, 2011, to properly account for endogeneity bias and nonlinear covariables effects. Secondly, we use a Weighted Bootstrap resampling procedure (Chatterjee and Bose, 2005, Chatterjee and Bose, 2000 and Bose and Chatterjee, 2002) to flexibly construct valid confidence intervals for the CS effect, in the presence of clustered observations at school level. Finally, the third innovation consists on applying an outlier identification method to avoid the bias generated by schools showing outlier values in class size. Such outlier detection method is similar to the procedure described in Dehon, Desbordes, and Verardi, 2015 for the standard instrumental variable estimator. The benefits of correctly handling such outliers is illustrated by means of a Monte Carlo simulation exercise.

We illustrate the proposed procedure by estimating the CS effect on Literature test-scores, using data from sixth grades at primary school in Uruguay. As in Angrist and Lavy, 1999, we exploit educational regulations in Uruguay that impose a ceiling of 40 students per class but our analysis is performed employing student-level micro data, instead of class-level aggregated data.

The Chapter is organized as follows. Section 3.2 describes the sources of class size endogeneity and the identification strategy for CS effect. Section 3.3 defines alternative estimation and inference methodologies. The data set to be used and descriptive statistics of relevant variables are presented in Section 3.4. Empirical results are reported in Section 3.5. Section 3.6 presents the simulation results illustrating the bias induced by outlier class size values and the performance of the proposed trimming strategy. Finally, the main conclusions are discussed in Section 3.7.

3.2 Class size endogeneity and the identification strategy

Estimating class size effect is a difficult task because of the endogenous nature of CS. Several sources of such endogeneity can be present simultaneously, even when analyzing a unique population of schools. Plausible sources are the following:

- One source derives from the potentially non-random assignment of students to classes, performed by school administrators. For example, if well-behaved students are allocated in bigger groups and the others in smaller ones, and student behavior is correlated with quality of learning process, then the CS effect (estimated without accounting for the endogeneity problem) will be upward biased. This mechanism can be derived from Lazaer's disruption model mentioned at the introduction (Lazear, 2001).

- A second source of endogeneity arises because CS is determined by enrollment (i.e. the quantity of students enrolled at a specific school). High enrollment schools tend to present larger CS and are commonly located in high populated cities and towns. These kind of high populated locations tend to have families with better socioeconomic status than low populated ones. Such socioeconomic family status affects both student ability and stimulus to reach educational goals and is partially unobserved by the researcher, thus generating a source of unobserved confounders. Therefore, in this situation it is likely that estimators will have an upward bias. This potential endogeneity factor can be limited by using sub-samples of schools located in regions with similar socioeconomic indicators.

- A third source of endogeneity is that student's parents have the possibility to choose the school they prefer, between schools located near to the area they live. Typically, parents that are worried about making a *good choice* of school, are generally more educated, more informed and take more care about their children's education (some of these parents' characteristics are unobserved to the researcher). These unobserved characteristics are positively related to student achievement. Usually, this kind of parents (concerned with their children's education) selects a school based on knowledge about the school's quality (another unobserved factor for the researcher). Then, it is relatively more likely that there will be: (i) low-enrollment (low-quality) schools with relatively small CS; and (ii) high-enrollment (high-quality) schools with relatively large CS. This will produce an upward bias on estimations of CS effect. Alternatively, parents can simply choose schools with smaller classes, hoping that they are better than larger ones, and this will generate a downward bias on the estimator.

To deal whit the endogeneity in CS we rely on instrumental variables analysis, exploiting a national regulation rule that imposes a ceiling of 40 pupils per class. Our instrument is the same as the one proposed by Angrist and Lavy, 1999 and can be defined as follows.

Let e_s be enrollment in school s and PCS_{sc} the *predicted class size* of school s and class c (i.e. our instrumental variable). Assuming that enrolled sixth grade pupils are divided into classes of equal size, we have

$$PCS_{sc} = \frac{e_s}{int[(e_s - 1)/40] + 1},$$

where $int(n)$ is the largest integer lower or equal to n.

Predicted class size (PCS) represents the rule that schools facing an enrollment size less or equal to 40 must have only one class. Similarly, schools whit enrollment between 41 and 80 must accommodate students in two classes, and so on.

Figure 3.1 shows the relationship of enrollment with both PCS (dashed line) and the average class size in the actual dataset (continuous line).[1] Dotted horizontal lines at 20, 27 and 40 pupils

[1] Such actual average CS was computed running a nonparametric regression of observed CS on observed enrollment counts. Nonparametric regression was based on penalized spline estimation and was performed with the R (R Core Team, 2014) package "mgcv" (Wood, 2011). Regression had been intentionally infra-smoothed, aiming to capture deep local behavior of mean class size.

indicate the levels where *PSC* function has corners and sharp discontinuities. It can be seen in the figure that average CS tends to follow a similar pattern than *PCS* rule, showing an abrupt decreasing around discontinuities, but showing a smoother behavior. Detected discrepancies between the two lines are due to violations in the *PCS* rule which will be penalized by using the instrumental variable methodology.

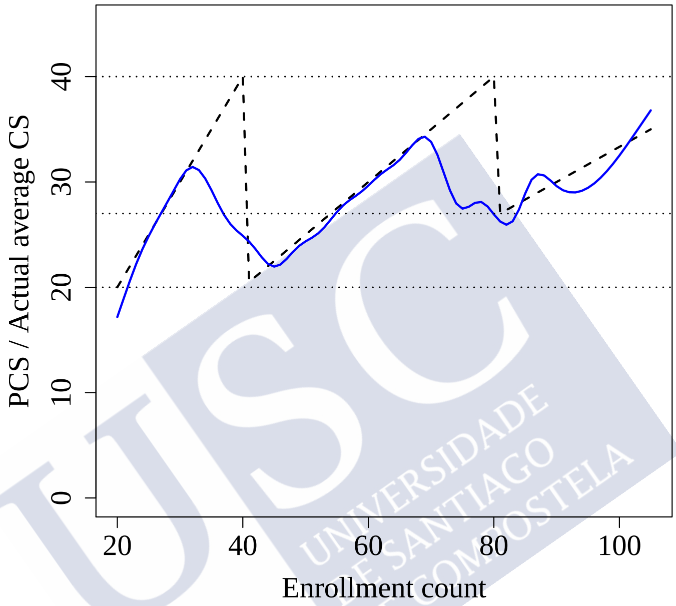

FIGURE 3.1: *Predicted class size and actual average class size as functions of enrollment count.*

In principle, the assignment rule given by *PCS* function (defined as a deterministic function of enrollment) can be seen as inducing an exogenous variation in CS. Using *PCS* as an IV implies leaving aside (i.e. penalizing) a portion of CS variability that is determined discretionally (i.e. endogenously) by schools administrators. In particular, for situations in which enrolled pupils are distributed into different sized classes based on their behavior or performance. Is important to realize here that *PCS* always predicts equal class sizes, within schools that need two or more classes, then it penalizes schools with asymmetric class sizes.

Additionally, this procedure helps correcting the bias caused by the parents' choice of school, whether the choice is made based on school quality or class size. If that choice is made considering school quality, then high-enrollment (high-quality) schools with large CS will tend to exceed *PCS*. On the contrary, low-enrollment (low-quality) schools with small CS will tend to be lower than *PCS*. Such deviations may arise, respectively, due to the lack or excess of schools's infrastructure (i.e. number of available classrooms). On the other hand, if parents act trying to get their children into a small class size, they would prefer schools with enrollment slightly larger than 40 or 80 pupils (i.e. hoping to fall on near the right side of a discontinuity). Because final enrollment is a random variable, parents are unable to predict class size with precision. Consequently, they can fall on with similar probability (i.e. randomly) at any side of discontinuities, avoiding the bias problem.[2]

Finally, enrollment itself is a possible source causing CS endogeneity, because it can be related with unobserved confounders affecting test score. Then *PCS* would partially include those confounders, affecting test scores by channels other than change in class size.

Assuming that such a confounders (and enrollment) have a smooth effect over test scores the problem can be mitigated because *PCS* breaks two times the enrollment path, allowing sharp discontinuities at 40 and 80 pupils. Therefore, the *PCS* function generates a large variability over CS at the discontinuities that mainly affect schools with *similar* enrollment (i.e. schools that stay near the discontinuities). But there are still schools that stay far from discontinuities. In principle, those schools have a reduced range of enrollment values comparing to the full range of enrollment, which reduces the potential bias effect caused by such variable. Therefore, to ensure unbiasedness it is necessary to introduce enrollment as a control variable in the model, in an attempt to capture all the remaining smooth effects from unobserved confounders. Following Angrist and Lavy, 1999 we use enrollment as an additional regressor with the purpose to get identification of class size effects.

3.3 Estimation and inference methodologies

3.3.1 Classical linear model approach

As described in Section 3.2 identification of CS effect has to be conducted using instrumental variables (IV) estimators. In the linear model framework, IV estimation may be performed via Two-Step Least Squares (2SLS), Generalized Methods of Moments for IV (GMM-IV) or Limited Information Maximum Likelihood (LIML), see for example Baum, Schaffer, and Stillman, 2007 and Angrist and Pischke, 2009 for a practical exposition. The three alternatives are equivalent in our case study, with one endogenous regressor and only one instrumental variable (i.e. the 'just identified" case).

Regarding inference procedure, it is crucial to consider a possible cluster dependence structure in the data. Test scores of students included in the same class or school may tend to be similar. This situation leads to a cluster dependence on the error components of the regression model. One possible solution consists in the explicit modeling of such a cluster correlation through inclusion of "random effects" (Moulton, 1986 analyzed that situation for the Ordinary

[2] Last discussion relates with Fuzzy Regression Discontinuity Design methodology (FRDD), formalized by Hahn, Todd, and Klaauw, 2001 and discussed by Angrist and Lavy, 1999 for the CS effect case.

Least Squares estimator and Shore-Sheppard, 1996 extended it to the instrumental variable setting). In that case, the test score regression model can be defined as

$$y_{isc} = X_s'\beta_s + X_c'\beta_c + X_i'\beta_i + \alpha CS_{sc} + \delta_s + \eta_c + \varepsilon_{isc},\ i = 1, \ldots, n, \tag{3.1}$$

where y_{isc} represents the i_{th} student's score, X_s', X_c' and X_i' are, respectively, vectors of school, class and pupil characteristics, and CS_{sc} is class size of class c in school s. The terms δ_s and η_c are i.i.d. random components representing school s and class c effects.[3]

Possible limitations of model (3.1), with respects to cluster-robust inference, relate to the imposition of substantive structure on the error components' data generating process (DGP). Primary it assumes that random terms are *identically distributed* (i.e. have the same distribution across clusters). Also it imposes additive separability of random components which, given the assumption of independence across clusters, implies the additivity of variance components (i.e. terms like $\sigma^2_\delta + \sigma^2_\eta + \sigma^2_\varepsilon$ as diagonal elements of the error's variance-covariance matrix).

Another implication of model (3.1) is that individual observations within a cluster have an error constant correlation (for example, $corr(\delta_{is}, \delta_{js}) = \rho$ for $i \neq j$), and it is the same in all clusters.

The parametric assumptions in (3.1) mentioned before make this approach sensitive to misspecifications. On the other hand, if those assumptions about the DGP are true, this strategy will produce the most efficient estimates.

An alternative and more flexible approach is the use of a cluster-robust estimator for the errors' variance-covariance matrix (White, 2014, 135-136). In this case the model specification is given by

$$y_{isc} = X_s'\beta_s + X_c'\beta_c + X_i'\beta_i + \alpha CS_{sc} + \varepsilon_{isc},\ i = 1, \ldots, n, \tag{3.2}$$

where $\varepsilon_{isc} = g(\delta_s, \eta_c, \varepsilon_{isc})$ is a general function of all random components. Then, the errors' variance-covariance matrix has the following block-diagonal form

$$\Omega = \begin{pmatrix} \Sigma_1 & & & & 0 \\ & \ddots & & & \\ & & \Sigma_m & & \\ & & & \ddots & \\ 0 & & & & \Sigma_M \end{pmatrix}, \tag{3.3}$$

[3] This modeling strategy was used in Angrist and Lavy, 1999 but aggregating data at the class level and using school level clustering.

where $\Sigma_m = \varepsilon_m \varepsilon_m'$ is the intra-cluster error covariance matrix for cluster m, denoting by ε_m the vector of errors for that cluster.

The matrix given by (3.3) is used in a weighting scheme either by 2SLS, GMM-IV or LIML estimation methods. Such methods produce the same results in our case of a unique endogenous variable and a unique instrumental variable. Actually, this is the most commonly used strategy when dealing with clustered data in applied research (see Baum, Schaffer, and Stillman, 2003 for a brief exposition).

The previous variance correction method is asymptotically valid in terms of the number of clusters. In fact, it works very poorly when the number of clusters is small.

A third alternative inference procedure, which combines flexibility and good performance with few clusters, consists in using some re-sampling scheme. Recently, some variations of the Bootstrap was proposed for inference in the context of IV parametric estimators with both independent observations (see for example [Davidson and MacKinnon, 2008, 2010, 2014]) and clustered data (Finlay and Magnusson, 2014). Usual bootstrapping types are *pair bootstrap* (i.e. *nonparametric* bootstrap) and the *wild bootstrap*.

Here we use a specific smooth kind of *weighted bootstrap* that belongs to the *generalized bootstrap* benchmark (Chatterjee and Bose, 2005). Additional discussion of the used weighted bootstrap is left to next subsection, where flexible additive models are presented.

Finally, since class-level clusters are embedded into school-level clusters (i.e. they are nested clusters) is enough to perform estimation assuming schools as the unique relevant clusters. That strategy is justified because both our regressors, on the one hand, and the unobserved errors, on the other hand, are correlated for each school.

3.3.2 Flexible additive model approach

In this subsection we present a flexible additive model approach to IV-based estimation along with a flexible resampling-based inference strategy. The additive model structure enables us to estimate smooth nonlinear effects of continuous control variables.

The specific model consists in a triangular simultaneous equations system, recently developed by Marra and Radice, 2011 for the semiparametric additive case and first proposed by Newey, Powell, and Vella, 1999 for the full nonparametric case.

The model structure is given by the structural equation (3.4) and the reduced form equation (3.5),

$$y_{isc} = \sum_{j=1}^{J} f^j(X_s^j) + \sum_{k=J+1}^{K} f^k(X_c^k) + \alpha CS_{sc} + \varepsilon_{isc},\ i = 1,\dots,n, \tag{3.4}$$

$$CS_{sc} = \sum_{j=1}^{J} g^{j}(X_{s}^{j}) + \sum_{k=J+1}^{K} g^{k}(X_{c}^{k}) + \pi PCS_{sc} + u_{isc},\ i = 1,\dots,n, \tag{3.5}$$

where $f(x)$ and $g(x)$ are flexible smooth functions of regressor x, J is the number of covariates at school level and $K-J$ is the number of covariates at the class level. We omit notation on both, regressors at the student level and controls with parametric effects, because they are not significant and unnecessary, respectively, in our empirical application.

A point to be noted is that CS presents a constant marginal effect (α) instead of a smooth nonlinear effect. This is so because CS' effect is estimated relaying on the two jumps or discontinuities in *PCS*, described in Figure 3.1. In fact, the estimated CS effect emerges partially by comparing large classes (near to 40 pupils, on the left side of both jumps) against small ones (near to 20 pupils, on the right side of the first jump, and 27 students, on the right side of the second discontinuity). Then, this estimated effect incorporates the change of test-scores due to large changes in CS (contrary to small marginal changes). The natural way to marginalize the effect of such large change in CS is using a linear effect (α), which implies obtaining an average/global marginal effect (i.e. assigning the same effect for each additional pupil in the class). Therefore, specifying a nonlinear effect for CS could be questionable in this case.

Estimation is performed as a two stage procedure named *Two Stages Generalized Additive Model* (2SGAM, Marra and Radice, 2011), which we will call simply *Two Stages Additive Model* (2SAM).

The regression function (3.5) is adjusted in the first stage, obtaining the estimated reduced form equation (3.6),

$$CS_{sc} = \sum_{j=1}^{J} \hat{g}^{j}(X_{s}^{j}) + \sum_{k=J+1}^{K} \hat{g}^{k}(X_{c}^{k}) + \hat{\pi} PCS_{sc} + \hat{u}_{isc},\ i = 1,\dots,n, \tag{3.6}$$

from which the IV's strength can be evaluated and the residuals $\hat{u}$ are obtained.

In the second stage, the structural equation (3.4) is adjusted including the vector of first stage residuals $\hat{u}$ as a regressor to obtain (3.7) and the estimated CS effect $\hat{\alpha}$.

$$\hat{y}_{isc} = \sum_{j=1}^{J} \hat{f}^{j}(X_{s}^{j}) + \sum_{k=J+1}^{K} \hat{f}^{k}(X_{c}^{k}) + \hat{\alpha} CS_{sc} + \hat{f}_{\hat{u}}(\hat{u}_{isc}),\ i = 1,\dots,n, \tag{3.7}$$

In principle, any method for adjusting additive models can be used, but following Marra and Radice, 2011 we use penalized splines (p-splines), which is described in more detail in Section 5.2 (Appendix of Chapter 3).

Inference in the 2SAM context relies on Bayesian confidence bands and hypothesis tests (Wood, 2006b; Marra and Wood, 2012; Wood, 2013), originally proposed for the Generalized Additive Model (GAM) setting (i.e. with only one regression equation). This approach consists in simulation from the estimated posterior distribution of model coefficients. That procedure was extended by Marra and Radice, 2011 to be applied in the 2SGAM benchmark (i.e. with two or more triangular simultaneous equations).

Such Bayesian methodology is valid for random sampling scenarios (i.e. independent sample observations) and it has not yet been developed to deal with clustered observations. Then, Bayesian inference procedure is not straightforwardly applicable in the current empirical case study.

To perform valid inference in the presence of clustered data we propose the use of a specific type of smooth weighted bootstrap (Chatterjee and Bose, 2005, Chatterjee and Bose, 2000 and Bose and Chatterjee, 2002). This type of resampling includes other variations previously studied (see for example Newton and Raftery, 1994, Barbe and Bertail, 1995 and Rubin, 1981). This kind of bootstrap was recently studied for models with features similar to those we are facing (i.e. semiparametric equations forming a triangular system, estimated by penalized M-estimators using cluster-dependent data). For example, Ma and Kosorok, 2005 established its validity for semiparametric M-estimators and some kind of Penalized M-estimators, Chen and Pouzo, 2009 shows its validity for penalized semiparametric estimation of a conditional moment with nonparametric endogeneity, Cheng, Yu, and Huang, 2013 applied it to Generalized Estimating Equations for cluster-dependent data and Chernozhukov, Fernández-Val, and Kowalski, 2014 demonstrates its pertinence for triangular simultaneous equations estimated by quantile regression.

The proposed weighted bootstrap (WB hereafter) algorithm consists on drawing i.i.d. random positive weights $\{w_s\}_{s=1}^{S}$, each of them being assigned to students within the same cluster/school s (note that S is the total number of clusters). Therefore, every school of the working sample receives one specific random weight, generating one bootstrap sample. Then, 2SAM is performed for each bootstrap sample which consists in the weighted original data using $\sqrt{w_s}$ as weights. The key feature of WB is that it does not really perform resampling of observations or clusters. Instead it generates estimators' variability through perturbation of estimating equations by means of simulated random weights (Chatterjee and Bose, 2005).

The WB helps overcome a known drawback of classical nonparametric bootstrap (or pair bootstrap), related to an over-smoothing tendency when nonparametric curve estimators are applied to the bootstrap samples. That problem emerges because repeated observations may appear in the generated samples. This situation is probably exacerbated in our study case, where entire clusters need to be re-sampled, leaving to repeated clusters at the generated samples. Instead, using WB ensures that bootstrap clusters are always different from each other.

The algorithm of the bootstrap methodology mentioned above can be sketched as follows.

Repeat B times steps 1 to 4 to get a number of B bootstrap estimations of the CS effect $\{\hat{\alpha}_b\}_{b=1}^{B}$:

1. Draws S random weights from a distribution with $\mathrm{E}(w_s) = 1$ and $\mathrm{Var}(w_s) = 1$). Following the recommendations in Chatterjee and Bose, 2005 and Chernozhukov, Fernández-Val, and Kowalski, 2014, we use weights obtained from a standard exponential distribution, $w_s \sim \text{Exponential}(1)$.

2. Assign each weight w_s to each school s, with $s = 1, \ldots, S$.

3. Weight each student data point at school s using the square root of its corresponding school weight ($\sqrt{w_s}$, with $s = 1, \ldots, S$).

4. Perform 2SAM over the weighted dataset and save estimated coefficient $\hat{\alpha}_b$.

Once bootstrap values $\{\hat{\alpha}_b\}_{b=1}^{B}$ have been obtained, estimating the variance and constructing the percentile confidence intervals is straightforward.

This WB algorithm was applied to make inferences about all parametric components (as our parameter of interest $\hat{\alpha}$) of the estimated models in Section 3.5., as justified by bootstrap validity results in Ma and Kosorok, 2005.

3.4 Data and descriptive statistics

The available data set comes from a national testing program for elementary schools in Uruguay. The program covers complete school population and was implemented during 1996. Data includes a standardized evaluation test, for literature and mathematics, conducted with sixth grade students. Students' scores in literature are used to approximate educational achievement. The score originally ranges between 0 and 20 but was re-scaled to take values between 0 and 100.

To obtain a relevant measure of class size we count the number of students who took at least one of the two tests. This CS measure was preferred over an alternative measure indicating maximum CS, because better approximates the size of the class through the complete scholar year. For 96% of the working sample, differences between the two class size measures are less or equal to 3 students. In any case, results remain fairly the same if the alternative measure is employed.

Data set also contains additional information about students, teachers and schools characteristics, enabling us to incorporate several control variables at the student and school level.

Because our main purpose is to illustrate the relevance of the methodological innovations mentioned in the introduction, we have restricted our analysis to a specific sub-sample of the dataset. The guiding idea was to apply our inference procedure to a specific sub-population of classes which presents a statistically and educationally significant CS effect. That idea leaded us to consider the afternoon shift classes in Uruguay's main department (i.e. Montevideo), excluding rural schools, full time schools, and schools classified as social disadvantaged. Using this restricted sample of classes is aligned to usual research practice which tend to study the class size effect at specific sub-populations, due to the substantive heterogeneity of such effect. In any case, the qualitative conclusions about the relevance of innovations suggested in the present work remain the same if bigger sub-samples are used.

The first working sample we analyzed is composed by 112 schools which include 187 classes and 4744 students. A second working sample that we considered includes 97 schools with 159 classes and 4111 pupils. This second sample excludes 28 outlier classes that severely violate the students allocation rule implied by PCS_{sc}, which was detected applying a similar strategy than proposed by Dehon, Desbordes, and Verardi, 2015. Additional details on the outliers detection method are presented in sub-section 3.5.2.

Variables included in the analysis are listed below:

- Score Literature: the student score in a literature test, including 20 questions, normalized between 0 and 100. This is the dependent variable in structural equations 3.2 and 3.4.
- Class Size (CS): the number of pupils per class (aproximated by the number of students who took at least one of the two evaluation tests).
- Predicted class size (PCS): it is the instrumental variable defined by (3.1).
- Enrollment: total number of students enrolled at school. It counts the students that effectively appear on classes lists.
- Socioeconomic index: a school level index characterizing *economic context* of the area where school is located, which takes a smaller value when the location is more disadvantaged.
- High Education (%): class level percentage of students having at least one parent with university studies.
- Housing issues (%): class level percentage of students with housing problems, defined as having more than two people per room, on average.
- Repeaters (%): Class level percentage of repeating pupils.

Additional student level variables were available but their inclusion as controls did not affect the CS effect's estimated magnitude nor its precision.

Table 3.1 presents descriptive statistics for variables of interest from the dataset that includes outliers. The literature scores distribution is centered around 64.18 and shows a moderate dispersion. Average class size is about 28 pupils, ten percent of classes have fewer than 22 students and ten percent have more than 34 pupils. Predicted class size presents a distribution similar to CS, with slightly bigger values at each quantile, as expected. Average enrollment is about 59 students. On average, classes have 33%, 20% and 28% of students with at least one parent with university education, having housing problems and being repeaters, respectively.

TABLE 3.1: *Descriptive Statistics (using sample with outliers).*

				Quantiles					
Variable	Mean	s.d.	min.	0.10	0.25	0.50	0.75	0.90	max.
Score Literature	64.2	18.7	4.2	37.5	50	66.7	79.2	87.5	100
Class Size (CS)	27.8	4.8	14	22	24	28	31	34	41
Predicted class size	29.9	5.3	20.5	22.5	26	29	33.3	38	40
Enrollment	58.6	16.6	21	36	48	58	71	79	103
Socioeconomic index	9.3	29.6	-58.3	-30.7	-11.1	11.1	30.9	50	70.1
High Education (%)	32.9	20.7	3.7	8	16.7	27.6	48.3	62.9	92
Housing issues (%)	20.3	13.7	0	4.7	10.3	18.2	27.3	40.6	64.7
Repeaters (%)	28.3	15.6	3.4	9.4	17.2	26.1	38.5	50	85.7

Table 3.2 shows descriptive statistics from the second working sample (which excludes outlier schools). In this case we can appreciate that CS and *PCS* present more similar distributions (more alike in both central tendency and dispersion) than in the previous working sample. The remaining variables show similar distributions in both samples.

TABLE 3.2: *Descriptive statistics (in sample without outliers).*

				Quantiles					
Variable	Mean	s.d.	min.	0.10	0.25	0.50	0.75	0.90	max.
Score literature	64.8	18.7	4.2	37.5	50	66.7	79.2	87.5	100
Class Size (CS)	28.2	4.6	17	22	24	29	32	34	38
Predicted class size	28.8	4.5	20.5	22.5	25.5	28.7	32	35	38.5
Enrollment	56.9	16.2	21	32	47	57	67	83	103
Socioeconomic index	9.2	28.8	-58.3	-28.6	-12	9.67	31.8	48.3	68.6
High education (%)	32.8	19.9	3.7	9.7	17.2	28.1	48.1	62.5	80
Housing issues (%)	20.2	13.7	0	5.7	10.3	17.9	27.3	40.6	64.7
Repeaters (%)	28.1	15.3	3.4	10	17.2	25.8	38.2	50	85.7

Before moving to the next section, it is useful to illustrate here the effect of the outliers trimming strategy on the identification possibilities using instrumental variable *PCS*. Figure 3.2 replicates Figure 3.1 (i.e. predicted class size and actual average class size as functions of enrollment count) but adding an alternative estimation for actual average class size, obtained using the sample without outliers.

A great improvement in co-variation between PCS and actual average class size can be appreciated when outliers are excluded, implying that schools in the trimmed sample tend to follow more closely the enrollment rule *PCS* path. This fact increases the reliability of the identification strategy based on instrument *PSC* presented in Section 3.2.

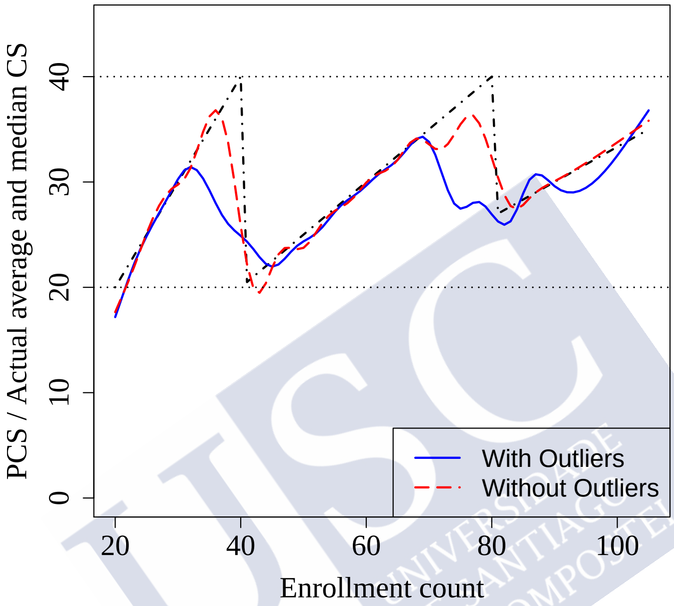

FIGURE 3.2: *Predicted class size and actual average class size as functions of enrollment count using both, full sample (continuous line) and sample excluding outlier schools (discontinuous line).*

3.5 Empirical results

3.5.1 Results ignoring endogeneity

If class size is considered an exogenous regressor (i.e. its potential endogeneity nature is neglected) then simple Ordinary Least Squares (OLS) estimates of parametric structural equation (3.2) would be enough to consistently estimate the CS effect α. Similarly, estimating the single flexible structural equation (3.4) would produce an estimation of the desired effect.

Tables 3.3 and 3.4 present OLS and REML (Restricted Maximum Likelihood) estimates of the test score regression equations (3.2) and (3.4), respectively, assuming CS exogeneity (i.e. without any endogeneity correction). The estimation was carried out using the larger sample consisting of 112 schools (that includes outlier schools). For the parametric model, we report usual (naive) standard errors, cluster-robust corrected standard errors (acording to variance-covariance matrix (3.3)), and clustered weighted bootstrap versions. As mentioned previously, correction was made assuming intra-school correlation. Additionally, the 95% and 99% weighted bootstrap confidence intervals (C.I.), based on bootstrap percentile method, are reported for each estimated parameter.

From both tables it can be concluded that neither OLS nor REML estimates provide evidence of a statistically significant CS effect on literature scores. In the OLS case, it can be seen that Enrollment, High education, Housing issues and Repeaters are statistically insignificant. Only Socioeconomic index at class level appear to have an appreciable effect on test scores.

TABLE 3.3: *OLS estimates of structural equation (including outliers)*

			Robust	Robust	Bootstrap percentile C.I.	
Regresor	coef.	s.e.	Asy. s.e.	WB s.e.	95% C.I.	99% C.I.
Class Size	0.0321	0.056	0.138	0.074	[-0.112, 0.177]	[-0.138, 0.229]
Enrollment	0.2619	0.0861	0.2164	0.113	[0.057, 0.489]	[-0.034, 0.548]
Enroll. squared	-0.0024	0.0007	0.0018	0.001	[-0.004, -0.001]	[-0.005, 400000]
Socioec. index	0.1610	0.0195	0.0539	0.028	[0.107, 0.216]	[0.089, 0.224]
High education	0.0588	0.0232	0.0602	0.031	[0.00022, 0.123]	[-0.023, 0.140]
Housing issues	-0.0389	0.0313	0.0644	0.033	[-0.102, 0.023]	[-0.120, 0.043]
Repeaters	-0.0857	0.0278	0.0599	0.031	[-0.146, -0.028]	[-0.169, -0.008]
Constant	56.42	2.994	7.259	3.827	[48.65, 63.58]	[46.42, 66.37]

TABLE 3.4: *REML estimates of structural equation (including outliers)*

			Robust	Bootstrap percentile C.I.	
Regresor	coef./(e.d.f)	s.e.	WB s.e.	95% C.I.	99% C.I.
Class Size	-0.0166	0.0589	0.066	[-0.138, 0.113]	[-0.173, 0.158]
Enrollment	(1.944)	-	-	-	-
Socioecon. index	(4.605)	-	-	-	-
High education	(4.590)	-	-	-	-
Housing issues	(4.118)	-	-	-	-
Repeaters	(1.001)	-	-	-	-
Constant	64.64	1.655	1.813	[60.92, 68.02]	[59.74, 68.81]

REML based estimation reveals the existence of smooth non-linear effects for Socioeconomic context, High education and Housing issues, which show empirical degrees of freedom (e.d.f) above 4. To illustrates that such non-linearities are complex enough to justify the flexible models used in this work, instead standard parametric alternatives (for example polynomial forms), Figure 3.3 shows graphically the flexibly estimated effects of covariates included in Table 3.4. In particular, effects of Socioeconomic Index and High Education present complex significant non-linearities that persist nearly with the same structure for the models that are estimated next in this chapter.

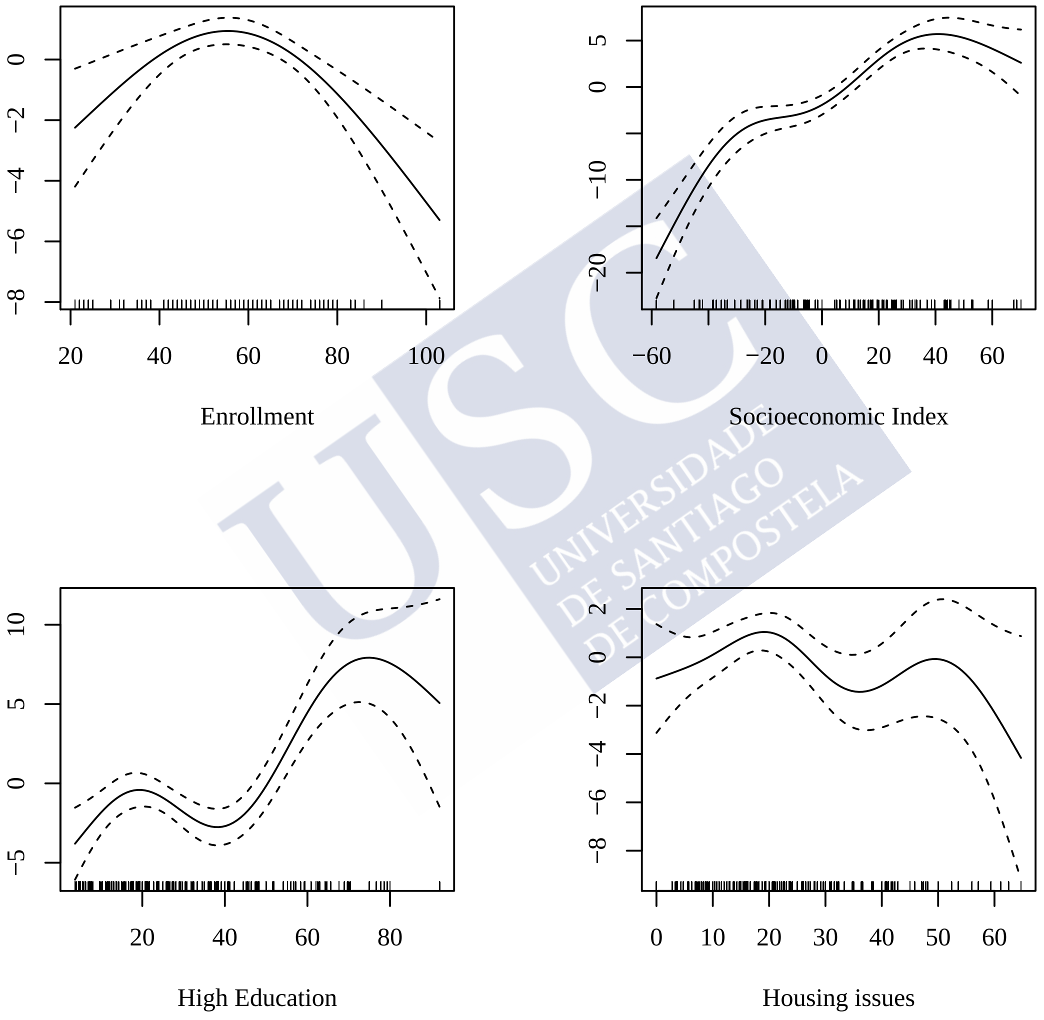

FIGURE 3.3: *Estimated flexible terms for Enrollment, Socioeconomic Index, High Education and Housing issues.*

Corrected standard errors are larger than naive ones. It must be noted that asymptotic robust standard errors are larger than their bootstrap versions. To judge which procedure is preferable in the current application we need to evaluate the characteristics of errors components' DGP imposed by asymptotic correction. Such DGP assumes constant correlation of errors components for all pairs of students within the same school. This assumption can be inadequate for schools with more than one class. Another feature to take into account relates to the number of clusters that the asymptotic correction requires to be valid. We have a moderate number of 112 clusters in our first working sample, which is not too small nor too large.

On the other hand, weighted bootstrap resampling appear to be more general or flexible than the asymptotic approach. First, it is valid even when the number of clusters is small (for example, empirical simulation results in Cheng, Yu, and Huang, 2013 used 25, 30 and 35 clusters). And second, it preserves the sample correlation of each pair of errors components in the same cluster, allowing non-constant correlation within clusters.

In addition to the arguments given above, our analysis relies on cluster bootstrap standard errors and confidence intervals, because of its validity in both parametric linear model and semiparametric additive model specifications. However we keep reporting statistics based on asymptotic correction for parametric models.

TABLE 3.5: *OLS estimates of structural equation (excluding outliers)*

			Robust	Robust	Bootstrap percentile C.I.	
Regresor	coef.	s.e.	Asy. s.e.	WB s.e.	95% C.I.	99% C.I.
Class Size	-0.1629	0.067	0.1758	0.092	[-0.337, 0.032]	[-0.393, 0.103]
Enrollment	0.2294	0.0869	0.2160	0.115	[-0.011, 0.458]	[-0.069, 0.539]
Enrollm. square	-0.0016	0.0007	0.0018	0.001	[-0.004, 0.001]	[-0.004, 0.001]
Socioecon. index	0.1977	0.0214	0.0597	0.031	[0.134, 0.257]	[0.119 , 0.277]
High education	0.0564	0.026	0.0670	0.036	[-0.012, 0.127]	[-0.046, 0.152]
Housing issues	-0.0052	0.0331	0.0684	0.036	[-0.074, 0.068]	[-102, 0.085]
Repeaters	-0.0529	0.0295	0.0656	0.033	[-0.120, 0.009]	[-0.139, 0.032]
Constant	59.79	3.157	7.873	4.145	[51.5, 67.6]	[48.9, 71.1]

TABLE 3.6: *REML estimates of structural equation (excluding outliers)*

			Robust	Bootstrap percentile C.I.	
Regresor	coef./(e.d.f)	s.e.	WB s.e.	95% C.I.	99% C.I.
Class Size	-0.2163	0.0589	0.087	[-0.380, -0.029]	[-0.454, 0.014]
Enrollment	(1.897)	-	-	-	-
Socioecon. index	(4.683)	-	-	-	-
High education	(4.530)	-	-	-	-
Housing issues	(4.233)	-	-	-	-
Repeaters	(1.006)	-	-	-	-
Constant	70.87	2.077	2.432	[65.58, 75.45]	[64.20, 77.14]

We repeated previous uncorrected OLS and REML estimation procedures but using the second working sample with 97 schools which excludes outliers. Tables 3.5 and 3.6 presents the corresponding results.

Exclusion of outlier schools increases the CS's negative estimated effect in both OLS (-0.163) and REML (-0.216) cases, but it still remains non significant with the exception of REML estimation using a bootstrap type 95% confidence interval. In the later case, CS effect ranges between -0.38 and -0.029, showing a lower limit staying very close to 0 from a practical perspective.

It is remarkable that REML based estimations show a larger negative CS effect relative to OLS counterpart. This is due to the existence of nonlinear effects for key control variables (i.e. Socioeconomic context, High education and Housing issues), which present empirical degrees of freedom (e.d.f.) with values above four.

3.5.2 Results correcting for endogeneity in the presence of outliers

This subsection presents IV based estimates of test score regression models (3.2) and (3.4) for both working samples (i.e. including and excluding outliers).

The following presentation is divided into three parts. In the first part attention is focused on the estimation of first stage equations and the influence of outlier schools. The second part presents IV based estimates of test score regression models (3.2) and (3.4) for both working samples. Finally, the third part describes the proposed procedure to detect outliers observations to be excluded.

We report cluster-robust standard errors based on weighted bootstrap (for all models) and asymptotic correction (for parametric models), both calculated taken schools as the relevant clusters.

First stage IV estimation

First stage reduced form equation, relating CS with the instrumental variable *PCS*, permits decomposing class size into an exogenous part (the systematic component of the equation) and an endogenous portion (the equation's random component). The endogenous part represents the *deviations* of CS from *PCS* rule, which can be seen as penalizing (at second stage) those schools that violate the rule.

From a theoretical perspective, if we assume a scenario without endogeneity then schools will tend to comply the assignment rule *PCS*. Therefore, in that scenario, it is expected to see *PCS* determining CS through a linear (constant) effect approximately equal to 1. Given this insight we expect to get an estimated *PCS* effect fairly close to 1, at the first stage regression, to be able to correctly decompose class size into its exogenous and endogenous parts.

It can be seen in Tables 3.7 and 3.8 that *PCS* presents an effect of about 0.35 on class size, which is a small value relative to the theoretical expected effect. This discrepancy is not produced by the additional regressors, in fact, a simple linear regression model relating CS exclusively to *PCS* produces an estimated effect of about 0.42. The left graph in Figure 3.4 shows the simple linear adjustment over scatter-plot of CS vs. *PCS*.

TABLE 3.7: *OLS estimates of reduced form equation (including outliers)*

			Robust	Robust	Bootstrap percentile C.I.	
Regresor	coef.	s.e.	Asy. s.e.	WB s.e.	95% C.I.	99% C.I.
Predicted CS	0.3554	0.0135	0.1184	0.061	[0.238, 0.477]	[0.188, 0.499]
Enrollment	0.1463	0.0208	0.0949	0.050	[0.053, 0.250]	[0.023, 0.296]
Enroll. squared	-0.0010	0.0002	0.0007	0.0004	[-0.002, -0.0003]	[-0.002, 0.0002]
Socioec. index	-0.0490	0.0046	0.0257	0.014	[-0.074, -0.021]	[-0.082, -0.014]
High education	0.0234	0.0056	0.0292	0.015	[-0.006, 0.052]	[-0.015, 0.060]
Housing issues	-0.0953	0.0074	0.0351	0.019	[-0.129, -0.058]	[-0.141, -0.044]
Repeaters	0.0041	0.0068	0.0342	0.018	[-0.031, 0.040]	[-0.046, 0.048]
Constant	13.94	0.7279	3.80	1.966	[10.16, 17.74]	[8.63, 18.62]
Bootstrap based F statistic for instrument *PCS*: 33.9						
Asymptotic based F statistic for instrument *PCS*: 9.01						

TABLE 3.8: *REML estimates of reduced form equation (including outliers)*

			Robust	Bootstrap percentile C.I.	
Regresor	coef./(e.d.f)	s.e.	WB s.e.	95% C.I.	99% C.I.
Predicted CS	0.3527	0.0133	0.059	[0.241, 0.472]	[0.193, 0.493]
Enrollment	(1.929)	-	-	-	-
Socioecon. index	(4.958)	-	-	-	-
High education	(4.954)	-	-	-	-
Housing issues	(4.338)	-	-	-	-
Repeaters	(4.928)	-	-	-	-
Constant	17.23	0.404	1.692	[13.83, 20.37]	[13.23, 21.80]
Bootstrap based F statistic for instrument *PCS*: 35.7					

The real cause of the large observed discrepancy is the presence of a group of outlier schools (represented by hollow grey circles in Figure 3.4 scatter-plots) that largely violates the *PCS* rule, generating an evident downward bias in *PCS* effect. Correcting such bias is crucial to get a first stage adjustment which effectively decompose CS into its exogenous and endogenous parts. This correction can be obtained using the second working sample that excludes the referred outlier schools, as appreciated in right panel of Figure 3.4 were the estimated linear effect rounds 0.96 (a value very close to 1, as expected).

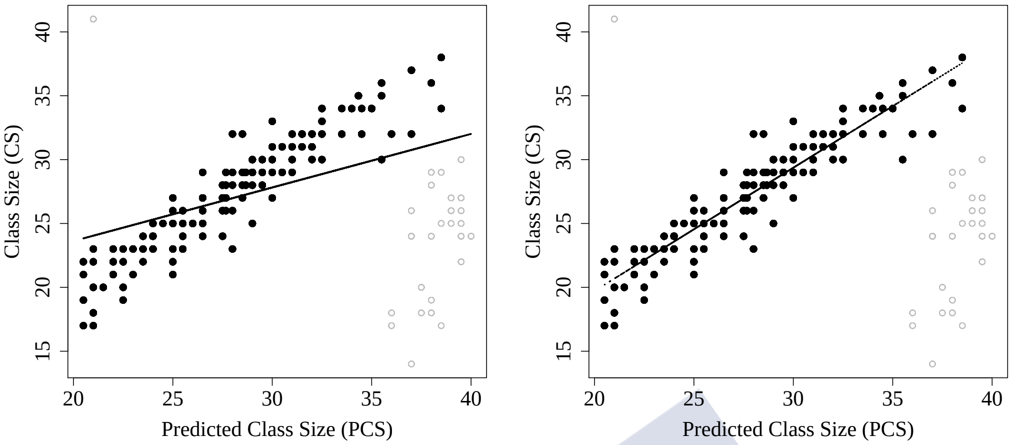

FIGURE 3.4: *OLS regression of CS on PCS using sample with outlier schools (left graph) and using sample excluding outliers (right graph).*

Tables 3.9 and 3.10 presents first stage regressions results, for Ordinary Least Squares (OLS) and Restricted Maximum Likelihood (REML) respectively, using the sample purged from schools.

TABLE 3.9: *OLS estimates of reduced form equation (excluding outliers)*

Regresor	coef.	s.e.	Robust Asy. s.e.	Robust WB s.e.	Bootstrap percentile C.I. 95% C.I.	 99% C.I.
Predicted CS	0.9481	0.0059	0.0262	0.014	[0.920, 0.974]	[0.912, 0.982]
Enrollment	-0.0197	0.0075	0.0263	0.014	[-0.046, 0.009]	[-0.053, 0.020]
Enroll. squared	0.0002	.00006	0.0002	0.0001	[-6.3e^{-05}, .0004]	[-.0001, .0004]
Socioec. index	-0.0008	0.0018	0.0093	0.005	[-0.010, 0.009]	[-0.014, 0.013]
High education	0.0128	0.0022	0.0093	0.005	[0.003, 0.022]	[-.0002, 0.025]
Housing issues	-0.0263	0.0028	0.0144	0.007	[-0.040, -0.012]	[-0.045, -0.007]
Repeaters	0.0198	0.0025	0.0178	0.009	[0.003, 0.038]	[-0.002, 0.046]
Constant	0.934	0.274	1.124	0.577	[-0.241, 1.961]	[-0.753, 2.324]
Bootstrap based F statistic for instrument *PCS*: 4586						
Asymptotic based F statistic for instrument *PCS*: 1309						

Both, OLS and REML give estimated values for the *PCS* effect very close to unity (0.948 and 0.919, respectively), as expected from theoretical considerations. These values are almost three times greater than their counterparts obtained using the full sample of schools, increasing the instrument *PCS* strength. For example, for REML estimation we get an F statistic (i.e. the concentration parameter) for *PCS* of about 35.7 using the full sample, but using the restricted sample we get an F statistic equal to 3754. Even though a concentration parameter value of 35.7

TABLE 3.10: *REML estimates of reduced form equation (excluding outliers)*

Regresor	coef./(e.d.f)	s.e.	Robust WB s.e.	Bootstrap percentile C.I. 95% C.I.	99% C.I.
Predicted CS	0.9191	0.006	0.015	[0.888, 0.948]	[0.879, 0.955]
Enrollment	(1.961)	-	-	-	-
Socioecon. index	(4.679)	-	-	-	-
High education	(4.848)	-	-	-	-
Housing issues	(4.701)	-	-	-	-
Repeaters	(4.881)	-	-	-	-
Constant	1.735	0.1750	0.424	[0.929, 2.579]	[0.719, 2.82]
Bootstrap based F statistic for instrument *PCS*: 3754					

excludes in principle a weak instrument scenario, increasing that value by more than 100 times constitutes a guaranty of a strong instrument.

A similar situation can be described in the OLS case. Using WB robust standard errors, F statistics including and excluding outliers are equal to 33.9 and 4586, respectively. And using asymptotic robust std. errors we get $F = 9.01$ with oultiers and $F = 1309$ without outliers. In the later situation, taking into account outlier schools makes a difference between a weak instrument and a strong instrument scenario.

Second stage IV estimation

Having described first stage estimation performance and confirmed the condition of PCS as a strong instrumental variable (except in the OLS case including outliers), now we focus our attention to the second stage estimation of the test scores structural equation.

Tables 3.11 and 3.12 report estimation results, based on the full sample (including outlier schools), of structural equations (3.2) and (3.4) respectively.

Both methods, Two Stage Least Squares (2SLS) and REML-based Two Stage Additive Model (2SAM), produce significant and similar point and interval estimations of the CS effect. The point estimations are very close to -1 and the interval estimations range between a minimum of -0.38 (2SAM's 99% C.I. lower limit) and -2 (2SLS's 99% C.I. upper limit).

These values suggest a large CS marginal effect on students test scores, relative to the values commonly founded in the literature.

The introduction of smooth effects for control variables, in the 2SAM procedure, did not produce a great impact on the CS effect magnitude nor in its estimation precision (relative to 2SLS), at least in this particular case.

TABLE 3.11: *2SLS estimates of structural equation (including outliers)*

Regresor	coef.	s.e.	Robust Asy. s.e.	Robust WB s.e.	Bootstrap percentile C.I. 95% C.I.	 99% C.I.
Class Size	-0.952	0.1618	0.5090	0.285	[-1.636, -0.515]	[-2.029, -0.412]
Enrollment	0.4375	0.0927	0.2679	0.144	[0.178, 0.748]	[0.085, 0.891]
Enroll. squared	-0.0032	0.0007	0.0022	0.001	[-0.005, -0.001]	[-0.006, -.0004]
Socioec. index	0.1093	0.0215	0.0645	0.033	[0.042, 0.171]	[0.026, 0.188]
High education	0.0836	0.0243	0.0701	0.036	[0.014, 0.156]	[-0.004., 0.175]
Housing issues	-0.1395	0.0358	0.0927	0.049	[-0.238, -0.052]	[-0.279, -0.023]
Repeaters	-0.0804	0.0287	0.0670	0.035	[-0.152, -0.012]	[-0.176., 0.008]
Constant	77.92	4.52	12.99	6.987	[65.19, 93.50]	[62.07, 100.3]

Hausman test of endogeneity, i.e significance test for first stage residuals inclusion at second stage (based on asymptotic cluster-robust std. errors):
z = 3.32; p-value = 0.0009 (p-value assuming normality)

Bootstrap-based Hausman test of endogeneity, i.e significance test for first stage residuals inclusion at second stage (based on WB cluster-robust std. errors):
z = 3.99; p-value = 0.00006 (p-value assuming normality)

TABLE 3.12: *2SAM estimates of structural equation (including outliers)*

Regresor	coef./(e.d.f)	Robust WB s.e.	Bootstrap percentile C.I. 95% C.I.	 99% C.I.
Class Size	-0.935	0.272	[-1.611, -0.502]	[-1.902, -0.382]
Enrollment	(1.953)	-	-	-
Socioecon. index	(4.714)	-	-	-
High education	(4.697)	-	-	-
Housing issues	(4.672)	-	-	-
Repeaters	(1.002)	-	-	-
Constant	90.16	7.60	[78.1, 109.2]	[74.48, 116.7]

There are two additional features to be noted in the parametric 2SLS case. First, bootstrap standard errors are smaller than their asymptotic counterparts. In fact, based on asymptotic standard errors, the CS effect seems to be statistically non-significant at 95% of confidence. Second, Hausman test of endogeneity rejects the exogeneity hypothesis both, when based on asymptotics and with the bootstrap versions of standard errors.

Finally, Tables 3.13 and 3.14 present the corresponding estimation results when the outlier schools are excluded from the sample. The main novel result is a substantial reduction of estimated CS marginal effect. 2SLS produces an estimation of -0.215 and 2SAM results in an effect equal to -0.314.

In this case there are some noticeable differences between the parametric and flexible methods. First, the point estimation of the CS effect using 2SAM is about 46% larger than its 2SLS counterpart. Second, the statistical significance of the CS effect is weaker in the parametric case (only slightly significant relying on 95% C.I.) than in the additive case (significant with both

TABLE 3.13: *2SLS estimates of structural equation (excluding outliers)*

			Robust	Robust	Bootstrap percentile C.I.	
Regresor	coef.	s.e.	Asy. s.e.	WB s.e.	95% C.I.	99% C.I.
Class Size	-0.2152	0.0726	0.1780	0.094	[-0.389, -0.013]	[-0.453, 0.042]
Enrollment	0.2364	0.0870	0.2164	0.116	[-0.009, 0.464]	[-0.070, 0.547]
Enroll. squared	-0.0016	0.0007	0.0018	0.001	[-0.004, .0004]	[-0.004, .001]
Socioec. index	0.1959	0.0214	0.0592	0.031	[0.133, 0.256]	[0.117, 0.273]
High education	0.0578	0.0260	0.0667	0.036	[-0.011, 0.127]	[-0.045., 0.155]
Housing issues	-0.0091	0.0332	0.0681	0.036	[-0.073, 0.066]	[-0.102, 0.083]
Repeaters	-0.0522	0.0295	0.0653	0.033	[-0.118, 0.011]	[-0.134., 0.033]
Constant	60.92	3.21	7.72	4.070	[52.85, 68.68]	[50.89, 72.21]

Hausman test of endogeneity, i.e significance test for first stage residuals inclusion at second stage (based on asymptotic cluster-robust std. errors):
z = 0.97; p-value = 0.332 (p-value assuming normality)

Bootstrap-based Hausman test of endogeneity, i.e significance test for first stage residuals inclusion at second stage (based on WB cluster-robust std. errors):
z = 1.88; p-value = 0.06 (p-value assuming normality)

TABLE 3.14: *2SAM estimates of structural equation (excluding outliers)*

		Robust	Bootstrap percentile C.I.	
Regresor	coef./(e.d.f)	WB s.e.	95% C.I.	99% C.I.
Class Size	-0.314	0.101	[-0.509, -0.113]	[-0.573, -0.067]
Enrollment	(1.783)	-	-	-
Socioecon. index	(4.691)	-	-	-
High education	(4.537)	-	-	-
Housing issues	(3.737)	-	-	-
Repeaters	(1.005)	-	-	-
Constant	73.63	2.87	[67.98, 79.22]	[66.71, 81.08]

95% and 99% C. I.'s).

Focusing on parametric estimation, the asymptotic approximation concludes that there is no evidence of an endogeneity problem, in light of the first Hausman test in Table 3.13. But the second Hausman test, based on weighted bootstrap, brings doubts about CS exogeneity showing a p-value of 0.06.

Outliers handling

Given the highlighted impact of outliers schools in the first and second stages, identification for this kind of observations become crucial to avoid a bias problem in the CS effect estimation.

To appropriately detect the relevant outliers affecting first-stage estimation, we propose application of a modified version of the strategy described in Dehon, Desbordes, and Verardi, 2015.

The standard detection method in Dehon, Desbordes, and Verardi, 2015 consist on two stages. In first stage, it identify outlying observations simultaneously in the response variable, the endogenous regressor, the instruments and the control variables, using the Stahel, 1981 and Donoho, 1982 (Stahel-Donoho hereafter) univariate projections estimator. In second stage, it apply the standard instrumental variable estimator to the outlier-free sub-sample.

Intuitivelly, the main idea of Stahel-Donoho method consist on transforming a multivariate outlier into a univariate one, projecting the data cloud in all possible directions to get a one-dimensional projection. Then, the degree of outlyingness of each data point is measured as the maximal univariate robust standardised distance from the centre of a given projection to that point.

In other terms, given a $(n \times k)$ dataset $\mathbf{x}$ (i.e with n observations and k variables), the outlyingness for each multivariate point x_i, relative to $\mathbf{x}$, is defined as its maximal univariate Stahel-Donoho outlyingness measured over all directions. To obtain the univariate Stahel-Donoho outlyingness in the direction d, the dataset $\mathbf{x}$ is projected on d, and the robustly standardized distance of $d'x_i$ to the robust center of the projected data points $\mathbf{x}d$ is computed.

There are several options to compute the mentioned robust standardised distance and to define the directions d. For example, the data for each projection can be centred around the median and standardised by the median absolute deviation, which is the alternative we employed.

The drawback of Dehon, Desbordes, and Verardi, 2015 methodology (DDV hereafter) is that, eventually, relevant bivariate outliers are masked as non-outliers, because they are not considered as outliers in a multivariate sense. This is the case for the bivariate distribution of CS and *PCS* in the first-stage regression.

It is important to note that the bivariate correlation between CS and *PCS* is very strong, therefore *PCS* explains a large part of the variability in CS at first-stage estimation. For this reason, all bivariate outliers which heavily affects the correlation between CS and *PCS* must be detected.

For the aforementioned arguments, we propose the application of DDV detection method but restricted on the bivariate distribution of CS and *PCS* (BDDV hereafter). The isolated focus on this two variable avoid the masking problem that occurs when all variables participating in first and second stages regressions are included in the outlyingness computation procedure.

Therefore, our proposed BDDV method consist on two stages. In first stage, we identify outliers in the bivariate distribution of CS and *PCS*, using the Stahel-Donoho univariate projections estimator. In second stage, we apply the standard instrumental variable estimator to the outlier-free sub-sample.

Figure 3.5 illustrates the mentioned masking problem, that arises when standard DDV detection method is applied, using the simple linear regression of CS on PCS at first stage. Comparing both scatter-plots in Figure 3.5 becomes evident that standard DDV procedure (left graph) fails to detect a group of bivariate outliers which are successfully detected by the proposed BDDV method. Such incomplete detection of standard DDV produces a downward-bias on the effect of PCS, which decreases from 0,96 to 0,63 when not detected outliers are present.

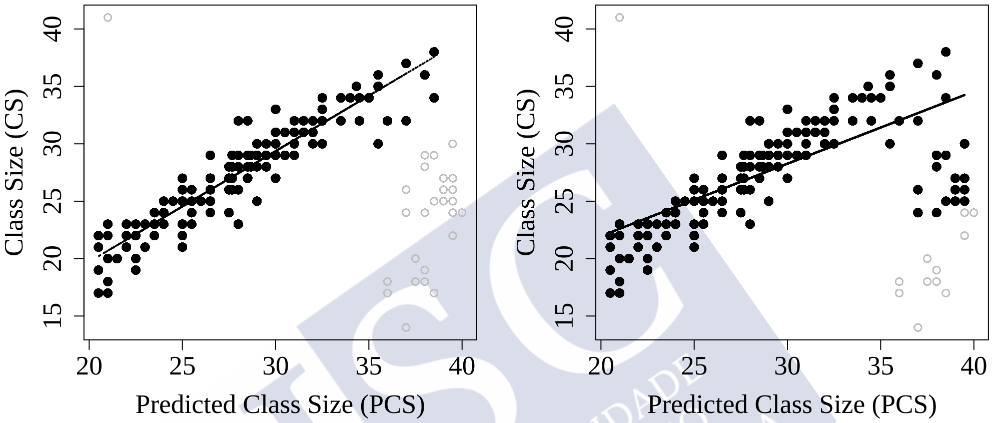

FIGURE 3.5: *OLS regression of CS on PCS using outlier-free sample based on proposed BDDV method (left graph) and using outlier-free sample based on standard DDV procedure (right graph).*

In a similar way, in the context of our flexible reduced form model (3.5), the estimated effect of PCS on CS decreases from 0,92 to 0,64 when standard DDV is applied instead of the BDDV version.

3.6 Monte Carlo simulation illustrating outliers impact

The problems generated by the existence of outliers observations in the context of IV estimation are well known in the specialized literature, see for example Dehon, Desbordes, and Verardi, 2015 and Zhelonkin, Genton, and Ronchetti, 2012 for recent treatments. Also known is that such a problems are often neglected or poorly handled in empirical research.

Through a simple simulation exercise we illustrates the estimation bias that can be generated by a portion of outliers observations in the 2SLS benchmark. Additionally, we show the effectiveness of the BDDV trimming strategy, that we proposed in previous section, to avoid the aforementioned bias.

The Data Generating Processes (DGP) designed, which is consistent with the endogeneity sources discussed previously and omit unnecessary complexities as control variables and clustered data structure, assumes the existence of two types or groups of classes. The first group is composed by classes at high quality schools, which present a higher expected enrollment and comply the assignment rule given by PCS (i.e. the official rule setting a maximum CS of 40 students). The second group includes low quality schools presenting a lower expected enrollment, which apply a different assignment rule setting a maximum CS of 30 students (henceforth PCS30). This group of schools is responsible for the outlier generation in CS values, with a similar pattern of outliers described in Figure 3.4. For these schools we assume that observed difference between the application of its own rule (PCS30) and the couterfactual official rule PCS (i.e $\triangle = PCS30 - PCS$), negatively affects the students achievement. This is consistent to the findings in our empirical dataset, in which the set of outlier classes present a lower mean test score (60.3 points) than the rest of classes (64.8 points), with that difference of 4.5 points being different from 0 with 99.9 % confidence level.

Additionally, DGP imply that low quality schools assigns well-behaved students in bigger classes and worst-behaved ones to smaller classes, in an attempt to improve scholastic achievement. This students allocation originates bigger classes with higher scores and smaller classes with lower scores. Finally, each observation simulated by the DGP represents a class-level aggregated information.

DGP for the high quality schools/classes is given by

$$
\begin{aligned}
&Enrollment_{hq} = 40 + \delta_{hq} \\
&PCS_{hq} = \frac{Enrollment_{hq}}{int[(Enrollment_{hq} - 1)/40] + 1} \\
&CS_{hq} = \pi_1 PCS_{hq} + \upsilon_{hq}, \\
&Score_{hq} = \alpha_0 + \alpha_1 CS_{hq} + \omega_{hq}, \\
&\text{ with } \omega_{hq} \sim N(0,20),\ \upsilon_{hq} \text{ taking integers in } [-5,2] \text{ with equal probability} \\
&\text{ and } \delta_{hq} \text{ taking integers in } [-25,25] \text{ with equal probability ,}
\end{aligned}
\tag{3.8}
$$

and DGP of low quality classes is established as follows

$$
\begin{aligned}
& Enrollment_{lq} = 30 + \delta_{lq} \\
& PCS30_{lq} = \frac{Enrollment_{lq}}{int[(Enrollment_{lq} - 1)/30] + 1} \\
& CS_{lq} = \pi_1 PCS30_{lq} + \eta + \upsilon_{lq}, \\
& PCS_{lq} = \frac{Enrollment_{lq}}{int[(Enrollment_{lq} - 1)/40] + 1} \\
& \triangle = PCS30_{lq} - PCS_{lq} \\
& Score_{lq} = \alpha_0 + \alpha_1 CS_{lq} + 2\eta + \triangle + \omega_{lq}, \\
& \text{ with } \omega_{lq} \sim N(0,20),\ \upsilon_{lq} \text{ taking integers in } [-3,1] \text{ with equal probability} \\
& \ \delta_{lq} \text{ taking integers in } [-18,18] \text{ with equal probability} \\
& \text{ and } \eta \text{ taking integers in } [-5,-4,-3,3,4,5] \text{ with equal probability .}
\end{aligned}
\tag{3.9}
$$

We set $\pi_1 = 1$, $\alpha_0 = 80$, CS effect $\alpha_1 = 0.5$ and the sample size $n = 200$ (100 high quality classes and 100 low quality ones). The role of class-level component η is to model the schools strategy of splitting students depending on its behaviour in classroom.

TABLE 3.15: *Bias, standard error and mean squared error of OLS and 2SLS estimators*

		Complete sample		BDDV-Trimmed sample		Complete sample
Regressor/Method		OLS	2SLS	OLS	2SLS	Outlier-Robust 2SLS
Class Size (CS)	*Bias*	0.595	-1.01	0.282	-0.003	-0.002
	Std.Error	0.088	0.352	0.064	0.079	0.102
	MSE	0.362	1.135	0.083	0.006	0.010
PCS (first stage)	*Bias*	-	-0.491	-	-0.017	-0.020
	Std.Error	-	0.07	-	0.038	0.056
	MSE	-	0.247	-	0.001	0.004

Percentage of classes considered as outliers: 13.5%
Number of simulated samples: 3000
Robust 2SLS was applied using the default options proposed by its authors in the R-library **riv**.

It can be seen from Table 3.15 that OLS estimator always present an upward bias for CS effect, however such bias is about 50% smaller when the trimmed sample is used. On the other hand, 2SLS method produces a large downward bias in CS effect using the complete sample and presents a negligible bias if trimmed sample is used. Additionally, in the last column we include the results obtained with a robust-to-outliers IV procedure recently proposed by Freue, Ortiz-Molina, and Zamar, 2013, which robustify the solution of the ordinary estimating equations of 2SLS. This robust IV estimator performs in a similar way than the ordinary 2SLS with the trimming strategy, except for showing a slightly larger standard errors for both CS and PCS effects.

The key fact behind the good performance estimating CS effect of both 2SLS with the trimmed sample and robust 2SLS, is the unbiased estimation of PCS effect in the first stage regression.

As expected, the sampling variability of all estimators is reduced when the outliers are correctly handled. The percentage of sample considered as outliers is 13.5% in the trimming strategy, a proportion similar to the one we have considered in our empirical illustration.

3.7 Main conclusions

Results in the present chapter demonstrate the relevance of concurrent methodological innovations we proposed in the context of standard IV estimation of the CS effect on students achievement. As an illustrative application we studied the class size effect on literature test scores, in a sample of schools in Montevideo (Uruguay).

Firstly, neglecting outlier schools' influence in the first stage IV estimation produces a large bias in the instrument (*PCS*)'s estimated effect. This bias causes a bias in the CS effect estimation in the second stage. Such bias represented, in our illustrative application, an overestimation of CS negative effect of about 342% in parametric case and 198% for the semiparametric additive model case (compared with estimates obtained by correcting the influence of outliers).

This finding is not exclusive of our analysis. In fact, analyzing available data from Angrist and Lavy, 1999 and taking account of potential outliers by our proposed detection procedure, we find a similar pattern of bias for both *PCS* and CS effects. For example, using outlier-free data of fifth grade, we find nonsignificant CS effects of -0.08 on reading test scores and of about -0.05 on math test scores. Those effects sizes are smaller than the values of -0.260 and -0,261 obtained with the full sample and reported for models (3) and (9) of Table IV in Angrist and Lavy, 1999. Moreover, when the detected outliers are dropped we find an estimated PCS effect equal to 0.89 in the first stage equation, which is larger than the value of 0.542 obtained with the complete sample; see model (2) of Table III in Angrist and Lavy, 1999. Such a downward bias in the first stage regression seems to be present in many of the recent articles on the subject. For example, Li and Konstantopoulos, 2016 obtain values of the PCS effect on CS between 0.1 and 0.5 for 10 of a total of 14 European countries and Gary-Bobo and Mahjoub, 2013 find PCS effects between 0.225 and 0.338 in four different grades at French junior high schools.

Our small simulation experiment helped to bring out the benefits of an adequate handling of outliers observations when 2SLS is used. The simulation results are qualitatively the same as those obtained in our empirical illustration and in re-analysis of Angrist and Lavy, 1999 data.

The general strategy of outliers exclusion can be criticized in terms of the loss of sample information concerning the relationship between CS and students achievement, specially when the share of outliers in sample is high. Alternatively, other strategies that avoid such a loss can be used relying on any outlier-robust estimator, in particular for the first stage regression.

Secondly, using a flexible additive model specification for the IV estimation helped isolate the CS effect, especially when outlier schools were excluded from the sample. The flexible specification helped to better account for smooth effects in key control variables, resulting in an estimated negative effect 46% larger with respect to the parametric model estimation.

In the third place, the implementation of flexible Weighted Bootstrap for inference was helpful in two ways. First, it enables us to perform cluster-robust inference, through standard errors and confidence interval computation, when flexible model specifications were used. Second, it proved to be an alternative to the asymptotic approximations both when the number of cluster is moderated and when parametric assumptions of asymptotic approach seem to be questionable.

In terms of empirical evidence, the combination of these innovations helped to identify a statistically and practically significant negative effect for class size. The effect's magnitude, -0.314, represents a relatively large effect in terms of previous findings reported by specialized literature.

Important statistical problems remain to be studied in the flexible model approach based on 2SAM estimator, mainly related to inference procedures. For example, construction of valid point-wise confidence intervals for smooth terms components and statistical tests to assess the exogeneity hypothesis when instrument presents a smooth effect, are of primary interest.

Chapter 4

Flexible estimation of Triangular Simultaneous Equations Models with Weak Instruments

4.1 Introduction

The wide family of regression estimators based on instrumental variables (IV) share a necessary identification condition, namely, the existence of 'sufficient' partial correlation between instruments and corresponding endogenous variables. Non-compliance of this condition is known as the problem of *weak instruments* or *weak identification*. In practical applications of standard IV estimators, this problem causes undesirable results, mostly related to significant finite-sample bias, loose of precision and unreliability of the asymptotic normality approximation.

The weak identification problem has been extensively studied over the past 20 years in the parametric regression context. Main contributions can be found in Bound, Jaeger, and Baker, 1995, Staiger and Stock, 1997, Stock and Wright, 2000, Kleibergen, 2002, Stock and Yogo, 2005, Newey and Windmeijer, 2009 and Andrews and Cheng, 2012, all of them belonging to the frequentist literature. Recently, bayesian approaches to IV estimation have been proposed, which perform better than traditional frequentist alternatives (in terms of bias and confidence interval coverage) in certain scenarios characterized by weak identification (see Burgess and Thompson, 2012 and Conley et al., 2008).

In the nonparametric regression context several efforts have been made in designing a reliable IV method for estimation and inference. One research lines was opened by Newey, Powell, and Vella, 1999 which proposes the use of nonparametric Triangular Simultaneous Equations Systems. This alternative involves a two stage estimation procedure, similar to Two Stages Least Squares (2SLS), and is generally known as the Control Function Approach.[1] It's main advantage is that it can exploit methodological advances from the nonparametric regression

[1] A different alternative to nonparametric IV regression is known as the *regularization approach*, were a regularization parameter is needed to solve an ill-posed inverse problem (see for example Newey and Powell, 2003, Darolles et al., 2011, Horowitz, 2014 and Shaw, Cohen, and Chen, 2016).

literature. The practical advantage results from the partition of the problem into two stages, each of them involving the estimation of single regression equations by standard nonparametric methods. More recent works implementing this flexible Control Function Approach are Pinkse, 2000, Su and Ullah, 2008 and Marra and Radice, 2011.

Despite these contributions to nonparametric IV estimation, the study of the weak instruments problem in this context has been largely neglected. One exception in the frequentist view is the work in progress provided by Han, 2014, which defines the weak identification problem for the flexible Triangular Simultaneous Equations Model, and proposes a penalized series estimation method that alleviates the weak instruments effect.

In this chapter we propose a new nonparametric bayesian IV method for Triangular Equations Models with one endogenous variable, presented recently by Wiesenfarth et al., 2014, which appear to be competitive alleviating the weak identification effect. This bayesian method has an advantage over the frequentist method presented in Han, 2014 as it performs the estimation of all needed tuning parameters (including smoothing parameters) from the available data. This advantage represents an invaluable benefit for applied research when nonparametric methods are used.

In this chapter we establish a performance comparison in the context of a weak identification scenario, between the later bayesian IV method and a convenient frequentist alternative known as Two Stages Generalized Additive Models (2SGAM) introduced by Marra and Radice, 2011. Both bayesian and frequentist alternatives are comparable in several aspects, e.g. automatic smoothing parameter estimation/selection, usage of splines for specification of the basis functions and full implementation through packages written in R (R Core Team, 2014).

Our final results provide contributions in two directions. First, it is shown that, when weak instruments are present, the bayesian estimator presents appreciable improvements (relative to the 2SGAM estimator) in terms of bias and efficiency. And second, the proposition established by Han, 2014 about the conditions under which an instrument must be considered as *nonparametrically* weak, is revised for the case of 2SGAM.

The remainder of the chapter is organized as follows. Section 4.2 discusses the main features of the weak identification problem in the context of the Control Function Approach to flexible IV estimation. Section 4.3 reviews the frequentist and the bayesian estimation methods to be compared (with more emphasis on the new bayesian estimator). Section 4.4 presents the Monte Carlo simulation results. Finally, the relevant findings are discussed in Section 4.5.

4.2 Weak identification in nonparametric IV estimation

This chapter is concerned with the weak identification effects in the context of flexible estimation of Triangular Simultaneous Equations Models. The analysis is restricted to the case of one endogenous variable in the *just identified* case (i.e. with only one corresponding instrumental variable). For simplicity, additional regressors are omitted without loss of generality. Then, the basic model can be defined as the following two equations system with additive error components

$$y_2 = f_2(y_1) + \varepsilon_2, \;\; y_1 = f_1(w) + \varepsilon_1 \tag{4.1}$$

where y_2 is a continuous response variable, y_1 is the (continuous) endogenous regressor, w is the instrument for y_1 and the random errors are represented by ε_2 and ε_1 (note that the usual additive constants, or intercepts, have been excluded to facilitates the exposition). The model flexibility comes from the fact that effects $f_2(\cdot)$ and $f_1(\cdot)$ are smooth functions with unknown functional form.

The first equation in (4.1) is known as the *structural equation*, and the second equation can be called as the *reduced form equation*. Given system (4.1), the interest resides in consistently estimating the expected value of y_2 conditional on y_1, i.e. the regression function (4.2).

$$E(y_2|y_1) = f_2(y_1). \tag{4.2}$$

In addition, we assume that there exists an endogeneity problem, i.e $E(\varepsilon_2|\varepsilon_1)$ is a non-constant function of ε_1. Under these conditions the usual estimation of $f_2(\cdot)$, applying any nonparametric estimator to the first equation in system (4.1), will be inconsistent.

The endogeneity problem can be mitigated if w is a valid instrument satisfying the identification assumptions given by

$$E(\varepsilon_1|w) = 0 \text{ and } E(\varepsilon_2|\varepsilon_1, w) = E(\varepsilon_2|\varepsilon_1), \tag{4.3}$$

from which it can be derived the following conditional expectation

$$E(y_2|y_1, w) = f_2(y_1) + E(\varepsilon_2|\varepsilon_1, w) = f_2(y_1) + E(\varepsilon_2|\varepsilon_1) = f_2(y_1) + f_3(\varepsilon_1), \tag{4.4}$$

where $f_3(\varepsilon_1)$, that is called the *control function* or *control variable*, represents a function of the error term from the second equation of system (4.1), i.e. the *reduced form* equation.

This identification result enables the use of a two stage procedure, the Control Function Approach mentioned in the previous section, to obtain a consistent estimator for $f_2(\cdot)$. In the first stage any nonparametric consistent estimator can be used to get $\hat{f}_1(\cdot)$ and obtain residuals $\hat{\varepsilon}_1 = y_1 - \hat{f}_1(w)$, which can be seen as consistent estimations of errors ε_1. In the second stage, an Additive Model (Hastie and Tibshirani, 1986) can be used to adjust the structural equation but including first stage residual $\hat{\varepsilon}_1$ as an additional regressor, obtaining:

$$y_2 = \hat{f}_2(y_1) + \hat{f}_3(\hat{\varepsilon}_1) + \hat{\varepsilon}_2, \tag{4.5}$$

where $\hat{f}_2$ is a consistent estimator for the effect of endogenous variable y_1. Consistency is obtained because once $\hat{\varepsilon}_1$ is controlled for, the remaining variability of y_1 results from the variation in w, which possesses an exogenous status due to identification assumptions (4.3).

The problem of weak identification arises when instrument w presents insufficient explanatory power for endogenous variable y_1 in the first stage regression. In the parametric linear IV regression model this possibility can be detected through a testing strategy studied by Stock and Yogo, 2005 for Two Stages Least Squares and Limited Maximum Likelihood estimators.

In the flexible nonparametric context, the weak identification problem was not seriously studied before a recent working paper, Han, 2014, that characterizes the weak IV problem as a concurvity issue. Concurvity arises in the structural equation (4.5) because endogenous variable y_1 tends to be equal to ε_1 when instrument w tends to be non-significant explaining y_1. More formally, this situation can be represented as

$$y_1 = f_{1,n}(w) + \varepsilon_1 \to \varepsilon_1 \ \text{ a.s. } \ \text{as } n \to \infty,$$

statement that is possible under a specific condition for weak identification (see pages 10 to 12 in Han, 2014 for technical details).

Representing the unknown functions to be estimated by means of a series of *basis functions*, the structural equation (4.5) can be expressed as

$$y_2 = \sum_{k=1}^{\infty} \hat{\beta}_{1k} b_k(y_1) + \sum_{k=1}^{\infty} \hat{\beta}_{2k} b_k(\hat{\varepsilon}_1) + \hat{\varepsilon}_2, \tag{4.6}$$

where the $b_k(.)$'s are the basis functions (e.g. B-splines, Fourier series or Legendre polynomials series).

Through representation (4.6), the mentioned concurvity issue becomes a more familiar multicollinearity problem as

$$b_k(y_1) \to b_k(\hat{\varepsilon}_1) \ \text{ a.s. } \ \forall k.$$

Due to this multicollinearity problem, estimators $\hat{\beta}_{1k}$ and $\hat{\beta}_{2k}$ are very unstable, even after truncation of the aproximating function series (i.e. $k \leq K \leq \infty$).

Han, 2014 proposes a penalized estimator to control the estimators' instability and then regularizing the weak instruments problem.

One of the main practical contributions in Han, 2014 is concerned with the characterization of *nonparametrically weak* instruments (i.e. if an instrument must be considered weak or strong when used in nonparametric estimation), which extends parametric model results established by Stock and Yogo, 2005 to the nonparametric case.

Assuming the following parametric linear model specification of the reduced form equation

$$y_1 = \pi_0 + \pi_1 w + \varepsilon_1, \tag{4.7}$$

it is possible to define an usual measure of strength for instrument w, called the *concentration parameter*, as

$$\gamma^2 = \frac{\pi_1^2 \sum_{i=1}^n w_i^2}{\sigma_{\varepsilon_1}^2}, \tag{4.8}$$

which coincides with the population version of the F-statistic to test global significance at the reduced form equation (4.7). In the parametric context, Stock and Yogo, 2005 established in 10 the concentration parameter's limiting value, so that for values smaller than 10 there exists a weak instrument problem. In Han, 2014 such a concentration parameter's threshold was settled at the larger value of 16, when a nonparametric model is estimated in the second stage.

The larger value of the concentration parameter was detected through the comparison, between the new penalized series estimator and its naive (unpenalized) version, over a sequence of specifications with increasingly weaker instrument. Specifically, the new threshold is set at a minimum so that, for values smaller than such minimum, the penalized estimator shows significantly better performance compared to the naive IV estimator.

The main drawback of Han, 2014 penalized estimator is the lack of an automatic procedure to perform the selection/estimation of both, the penalty parameter (to control estimator instability due to the weak instruments problem) and the smoothing parameters (to control smoothness of flexible model terms), in a simultaneous way.

Due to the empirical relevance of both the weak identification problem and the availability of an automatic procedure for tuning parameters selection, in the subsequent sections we present and compare two approaches for Additive Model estimation, which share the advantage of data-based estimation of smoothing parameters, but differs in its ability to alleviate the effects of weak IVs presence.

4.3 Description of the methods to be assessed

Over the past decade, empirical work based on nonparametric regression methods has increased its relevance, due in part to the increasing availability of computational implementation of algorithms for estimation and inference.

An important share of such an implemented estimators were freely available through specific packages written in R language (R Core Team, 2014). Nowadays, the R environment plays a key role inside the scientific community, providing implementation of new statistical methods, in general, and nonparametric regression estimators, in particular.

One of the most complete R packages presently available, designed to perform estimation of Additive Models and Generalized Additive Models, is the *mgcv* package (Wood, 2011, Wood, 2006a and Wood, 2004). It supports GAM estimation relying on alternative types of smoothing parameter selection methods (e.g. Generalized Cross Validation, Restricted Maximum Likelihood and Akaike Information Criterion) and splines basis functions (e.g. thin-plate regression splines, cubic regression splines and B-splines).

Using the *mgcv* routines, Marra and Radice, 2011 proposed an estimation methodology to fit Triangular Simultaneous Equations Systems through Two Stages Generalized Additive Models (2SGAM). This estimator, based on the Control Function Approach and described in the following subsection, is our chosen frequentist alternative to study the weak identification effects and their characterization.

On the other hand, the bayesian nonparametric IV regression estimator proposed by Wiesenfarth et al., 2014 is a competitive alternative to the later 2SGAM. This estimator, presented in subsection 4.3.2, is expected to possess certain advantages when the weak identification problem arises[2]. For that reason, such bayesian estimator constitutes our chosen alternative to carry out the comparison against 2SGAM in scenarios with weak instruments.

Both approaches, the frequentist and the bayesian, are based on the penalized splines (P-splines) concept, introduced by Eilers and Marx, 1996 and extended by Lang and Brezger, 2004 to the bayesian case (Bayesian P-splines).

4.3.1 Frequentist Additive Model approach

To introduce 2SGAM's estimation procedure in more detail, we start specifying the reduced form equation and the structural equation in an additive model format, where the data set of

[2] Such expectation is justified with a deeper analysis presented in section 3.2.

size n, $\{y_{2i}, y_{1i}, w_i\}_{i=1}^{n}$, is obtained by random sampling. The resulting triangular system of two additive equations is given by the following expressions

$$y_{1i} = \pi_0 + f_1(w_i) + \varepsilon_{1i},\ i = 1,\ldots,n \tag{4.9}$$

$$y_{2i} = \beta_0 + f_2(y_{1i}) + f_3(\hat{\varepsilon}_{1i}) + \varepsilon_{2i},\ i = 1,\ldots,n \tag{4.10}$$

where $f_3(\cdot)$ is the control function term, $\{\hat{\varepsilon}_{1i}\}_{i=1}^{n}$ are the residuals obtained by fitting the reduced form equation (4.9) and ε_{2i} are error components such that $E(\varepsilon_{2i}|y_{1i}) = 0$. The usual constant terms are represented by π_0 and β_0. In this context, the first and second stages of 2SGAM consist in estimating equations (4.9) and (4.10), respectively.

First, the reduced form equation (4.9) is estimated by minimizing the following objective function

$$\sum_{i=1}^{n}[y_{1i} - \pi_0 - f_1(w_i)]^2 + \lambda_1 \int_0^1 [f_1''(w)]^2 dw, \tag{4.11}$$

where the first term is the sum of squared errors (i.e. the traditional least squares objective) and the second term is the integrated square of the unknown function's second derivative. The last component of the objective function is introduced to penalize the *wiggliness* of term $f_1(\cdot)$ through the smoothing parameter λ_1, which controls the trade-off between the model's fit and model's smoothness (i.e. the bias-variance trade-off). Once the instrument's effect ($\hat{f}_1(\cdot)$) and the intercept ($\hat{\pi}_0$) have been obtained, the corresponding residuals $\hat{\varepsilon}_{1i} = y_{1i} - \hat{\pi}_0 - \hat{f}_1(w_i)$ can be obtained. These residuals are an input in the second stage estimation and represent that part of endogenous variable y_1 related to the error component of the structural equation (ε_2).

Next, equation (4.10) can be estimated by minimizing the following objective

$$\sum_{i=1}^{n}[y_{2i} - \beta_0 - f_2(y_{1i}) - f_3(\hat{\varepsilon}_{1i})]^2 + \lambda_2 \int_0^1 [f_2''(y_1)]^2 dy_1 + \lambda_3 \int_0^1 [f_3''(\hat{\varepsilon}_1)]^2 d\hat{\varepsilon}_1, \tag{4.12}$$

where λ_2 and λ_3 are the smoothing parameters of terms $f_2(\cdot)$ and $f_3(\cdot)$, respectively, and first stage residuals $\hat{\varepsilon}_{1i}$ are introduced as an additional regressor.

It is important to note in (4.12) that the estimation of the endogenous variable term $f_2(y_1)$ and the control function term $f_3(\hat{\varepsilon})$ is performed by optimizing the same objective function based on a global error criterion. This can lead to a confounded estimate of $f_2(y_1)$ due to inappropriate choices for the smoothing parameter λ_3, as is documented in Wiesenfarth et al., 2014.

This undesirable consequences can be exacerbated by the effects of the concurvity problem, in the special context of weak identification.

Estimation through objectives (4.11) and (4.12) can be performed by several algorithms implemented with the R package mgcv. Available algorithms differ in two principal aspects. First, there are alternative options to represent unknown smooth functions, depending on the basis functions used. We chose the mgcv package's default basis functions representation which relies on thin-plate regression splines (Wood, 2003), because they possess several advantages over other options (see chapter 4 of Wood, 2006a). Second, smoothing parameters estimation can be done using different approaches, including Generalized Cross Validation (GCV), Maximum Likelihood (ML), Restricted Maximum Likelihood (REML) and Akaike Information Criterion (AIC). GCV is the default method in the mgcv package, but REML has proved to be preferable in some contexts, being less prone to display local minima and offering some improvement in mean square error performance (Reiss and Todd Ogden, 2009 and Wood, 2011). Therefore, we used both the GCV and the REML approaches, obtaining some behavior differences between them, which are of interest beyond the main objectives of the Chapter.

4.3.2 Bayesian Additive Model approach

The bayesian nonparametric IV regression (BNIV) model can be defined by the following simultaneous equations system

$$\begin{gathered} y_{1i} = \pi_0 + f_1(w_i) + \varepsilon_{1i}, \quad y_{2i} = \beta_0 + f_2(y_{1i}) + \varepsilon_{2i},\ i = 1,\ldots,n. \\ \text{with } (\varepsilon_{1i}, \varepsilon_{2i}) \sim N(\mu_l, \Sigma_l),\ l = 1,\ldots,C. \end{gathered} \tag{4.13}$$

The main innovation in (4.13) consist in the flexible specification of the errors components distribution, $(\varepsilon_1, \varepsilon_2)$. It's assumed they follow a mixture of bivariate gaussian distributions. Therefore, errors components can be grouped into $C \leq n$ clusters, with means $\mu_l = (\mu_{1l}, \mu_{2l})^t$ and covariances

$$\Sigma_l = \begin{pmatrix} \sigma^2_{1l} & \sigma_{12,l} \\ \sigma_{21,l} & \sigma^2_{2,l} \end{pmatrix},\ l = 1,\ldots,C.$$

More specifically, the joint error distribution assumes an infinite mixture model with the following hierarchy:

$$
\begin{array}{rcl}
(\varepsilon_{1i}, \varepsilon_{2i}) & \text{i.i.d.} & \sum_{c=1}^{\infty} \phi_c \mathrm{N}(\mu_c, \Sigma_c) \\
(\mu_c, \Sigma_c) & \text{i.i.d.} & G_0 = \mathrm{N}(\mu | \mu_0, \tau_{\Sigma}^{-1} \Sigma) \mathrm{IW}(\Sigma | s_{\Sigma}, S_{\Sigma}) \\
\phi_c & = & v_c \left(1 - \sum_{j=1}^{c-1} (1 - \phi_j) \right) = v_c \prod_{j=1}^{c-1} (1 - v_j), \quad c = 1, 2, \ldots \\
v_c & \text{i.i.d.} & \mathrm{Be}(1, \alpha).
\end{array}
$$

In this specification, the mixture components (i.e. the clusters) are i.i.d. draws from the base measure G_0 (given by a normal-inverse Wishart distribution) of a Gaussian Dirichlet Process (DP) while the mixture weights are generated in a stick-breaking manner based on a Beta distribution depending on the parameter $\alpha > 0$ of the Dirichlet process. The concentration α determines the strength of belief in the base distribution G_0, which is the expectation of the Dirichlet process around which more mass will be concentrated for large α.

The expected number of components for a given sample size n is approximatively given by $\mathrm{E}(K^* | \alpha, n) \approx \alpha \log(1 + n/\alpha)$ (Antoniak, 1974). Thus, the parameter α is directly related to the number K^* of unique pairs (μ_l, Σ_l) in the data.

In order to avoid fixing K^* arbitrarily, α is estimated from the data and consequently this requires to set a new prior. The standard conjugate prior for α is a Gamma prior $\alpha \sim \mathrm{Ga}(a_\alpha, b_\alpha)$, with $a_\alpha = b_\alpha = 2$ as default choices. This allows both small and large values of α corresponding to many and few mixture components, respectively.

Since the model includes constants π_0 and β_0, it requires to ensure that $\mathrm{E}(\varepsilon_{1i}, \varepsilon_{2i}) = 0$ for identifiability. This is achieved by choosing $\mu_0 = (0,0)^t$ and constraining $\sum_{i=1}^{n} \mu_{1i} = \sum_{i=1}^{n} \mu_{2i} = 0$.

With respect to priors on the parameters in the base distribution G_0, the a diffuse gamma prior $\tau_\Sigma \sim \mathrm{Ga}(a_\Sigma/2, b_\Sigma/2)$ is established for τ_Σ, with default hyperparameters $a_\Sigma = 1$ and $b_\Sigma = 100$. On the other hand, although imposing an IW-prior on S_Σ is conceptually and computationally straight-forward, associated hyperparameter choice is unclear. Therefore, the default is set $s_\Sigma = 3$ obtaining $S_\Sigma = 0.2 I_2$ and thus $\sigma_{rl}^2 \sim \mathrm{IG}(1, 0.1)$ as a weakly informative prior on the residual variances (see Wiesenfarth et al., 2014 for technical justifications of hyperparameter choices).

Flexible effects specification

The definition of the unknown smooth terms $f_1(\cdot)$ and $f_2(\cdot)$ in model (4.13) is based on the Bayesian analogue to penalized splines (i.e. Bayesian P-Splines) as introduced by (Lang and Brezger, 2004). Thus, we assume that each of the smooth functions $f_r(x)$ (with $r = 1,2,$) of continuous covariate x can be approximated by a spline function $s_r(x)$ in the space of spline functions $S(d_r, \kappa_r)$ of degree d_r with knots $\kappa_r = \{x_{\min} < \kappa_1 < \kappa_2 < \ldots < \kappa_{K_r} < x_{\max}\}$, i.e. $s_r(x_r) \in S(d_r, \kappa_j)$. Since $S(d_r, \kappa_r)$ is a $(K_r + d_r + 1)$-dimensional vector space (a subspace of all d_r-times continuously differentiable functions), $s_r(x_r)$ can then be represented as a linear combination of suitable basis functions $B_{k_r}(x_r)$. Hence, smooth effects $f_1(\cdot)$ and $f_2(\cdot)$ can be expressed as

$$f_r(x) = \sum_{k=1}^{K_r+d_r+1} \beta_{rk} B_k(x) = X_r \beta_r, \text{ with } r = 1,2. \tag{4.14}$$

Due to their simplicity and numerical stability, B-spline basis functions are used.

Although the global smoothness properties are determined by the degree of the spline basis d_r, the variability of the resulting estimates depends on the location and number of knots. Instead of directly aiming at optimizing the number and position of the knots in a data-driven manner, the penalized spline approach relies on using a generous number of equidistant knots in combination with a penalty that avoids overfitting. The common rule of thumb is to choose $K_r = \min(n/4, 40)$.

In the frequentist framework, Eilers and Marx (1996) proposed to penalize the squared q-th order differences of adjacent basis coefficients, thereby approximating the integrated squared q-th derivative of the spline function. In the Bayesian framework, this corresponds to assigning a random walk prior to the spline coefficients to be estimated. Specifically, we use the second order random walk priors

$$\beta_{rk} = 2\beta_{r,k-1} - \beta_{r,k-2} + u_{rk}, \text{ with } u_{rk} \text{ i.i.d } N(0, \tau_r^2), \; r = 1,2; \tag{4.15}$$

which constitutes an explicit modeling for the second order difference of adjacent basis coefficients. The random walk variance τ_r^2 acts as an inverse smoothing parameter with small values corresponding to high smoothing, and with large values corresponding to a high variability of the estimated function. In the limiting case of $\tau_r^2 \to 0$, the estimated function approaches a a linear effect.

From the random walk specification, the joint prior distribution for the coefficient vector β_r can be derived as a partially improper multivariate Gaussian distribution with density

$$p(\beta_r|\tau_r^2) \propto \left(\frac{1}{2\tau_r^2}\right)^{\frac{rank(\Delta_r)}{2}} \exp\left(-\frac{1}{2\tau_r^2}\beta_r^t \Delta_r \beta_r\right)$$

where Δ_r is the penalty matrix given by the cross-product of a difference matrix D_r of appropriate order, i.e. $\Delta_r = D_r^t D_r$.

The bayesian prior specification is completed with a prior on τ_r^2, to include estimation of the variance and therefore to allow for a data-driven amount of smoothness. This prior consist in a conjugate inverse-gamma distribution with shape and scale parameters a_{τ_r} and b_{τ_r}, i.e. $\tau_r^2 \sim \mathrm{IG}(a_{\tau_r}, b_{\tau_r})$.

Finally, for parametric effects π_0 and β_0 , we use diffuse priors $p(\pi_0) \propto$ const and $p(\beta_0) \propto$ const, assuming a complete lack of prior knowledge.

The control function and smoothing parameters estimation

The BNIV model is closely related to the Control Function Approach, as can be seen considering the structural equation (4.16) derived from the system (4.13)

$$y_{2i} = f_2(y_{1i}) + E(\varepsilon_{2i}|\varepsilon_{1i}) + \varepsilon_i, \;\; \varepsilon_i \sim N(0, \sigma^2_{(2|1),l})$$
$$\text{with } \mathrm{E}(\varepsilon_{2i}|\varepsilon_{1i}) = \mu_{2l} + \frac{\sigma_{12,l}}{\sigma^2_{1,l}}(\varepsilon_{1i} - \mu_{1l}) \;\text{ and }\; \sigma^2_{(2|1),l} = \sigma^2_{2,l} - \frac{\sigma^2_{12,l}}{\sigma^2_{1,l}}, \tag{4.16}$$

where $E(\varepsilon_{2i}|\varepsilon_{1i})$ is the control function and $\sigma^2_{(2|1),l}$ is the conditional variance in cluster l, whit $l = 1, \ldots, C$. Note that these conditional moments and parameters come from the conditional distribution of mixtures of C bivariate normals. Thereby, mean and variance components may vary with i such that $\mathrm{E}(\varepsilon_{2i}|\varepsilon_{1i})$ and $(\varepsilon_1, \ldots, \varepsilon_n)$ may follow any functional form and distribution, respectively.

In contrast to the 2SGAM approach, in particular, and flexible frequentinst approaches, in general, in the BNIV model the control function term $E(\varepsilon_{2i}|\varepsilon_{1i})$ acts as a varying coefficient allowing the degree of endogeneity correction to be different over observations. Moreover, this control function is not a smooth function of ε_1, therefore it does not impose dependencies between the values it takes for adjacent errors ε_{1i}.

Control function approaches can be extremely sensitive to outliers in the error distribution, since they do not account for the high variability of the control function at extreme values of ε_1

where observations are scarce. BNIV has two features that helps to reduce the outliers effect. On the one hand, the non-constant variances σ^2_{1l} and means μ_{1l} shrink the error terms ε_{1i} toward their (non-constant) mean, reducing the weight of outlier errors. On the other hand, given that τ_Σ plays an important role for the smoothness of the error density, then a small τ_Σ allows the μ_{1i} to vary more strongly around its mean which translates in a possibly stronger downweighting of outliers in ε_{1i} depending on τ_Σ.

As described so far, the relevant smoothing parameters associated to the control function estimation are the number of mixture components (governed by the parameter α and the data) and parameter τ_Σ. It is important to note that these parameters are different from the smoothing parameters associated to estimation of terms $f_r(\cdot)$ (i.e. τ^2_r). This feature of BNIV plays a critical role because, in the control function approach, smoothing parameter choice is particularly delicate since smoothness of functions in the first stage and of the control function influence the way of endogeneity bias correction for $f_2(y_1)$.

The previously mentioned characteristic of BNIV contrast with the 2SGAM approach, which optimizes the same global error criterion to selects the smoothing parameter of both the control function and the effect of the endogenous regressor (as described in section 4.3.1).

In comparison with the frequentist approach, the distinctive features of BNIV seem to be useful in alleviating the effects of the concurvity problem in weak identification scenarios. In particular, its relative efficiency is high without a consequent relative bias compensation. Probably, such efficiency advantage is derived from, in first place, the robustness of control function $E(\varepsilon_{2i}|\varepsilon_{1i})$ to outlier values in ε_{1i} and, in second place, differentiating between, on the one hand, estimation of the smoothing parameters in the control function $E(\varepsilon_{2i}|\varepsilon_{1i})$ and, on the other hand, estimation of the smoothing parameters for terms $f_r(\cdot)$.

The estimation of the BNIV model is fully Bayesian, involving posterior means from Gibbs sampling steps in an efficient Markov Chain Monte Carlo (MCMC) implementation. This full Bayesian procedure can be performed using R package *bayesIV* (Wiesenfarth et al., 2014). It includes two alternative methods for estimation of the joint errors distribution through a Dirichlet process mixture. We use the default method in package bayesIV, based on implementation provided by R package DPpackage (Jara et al., 2011). The details on all full conditionals of the bayesian specification are given in the following.

Full conditionals

The full conditionals for the coefficients vector β_r of the smooth functions f_r (i.e. $r = 1, 2$ for equations in system (4.13)) are Gaussian

$$\beta_r|\cdot \sim \mathrm{N}(\mu_{\beta_r}, P^{-1}_{\beta_r})$$

with precision matrix

$$P_{\beta_r} = X_r^t \Sigma_{r|-r}^{-1} X_r + \frac{\Delta_r}{\tau_r^2},$$

where Δ_r is the penalty matrix of flexible effect f_r based on a random walk prior of second order and mean

$$\mu_{\beta_r} = P_{\beta_r}^{-1} X_r^t \Sigma_{r|-r}^{-1}(y_r - \tilde{\eta}_r - \mathrm{E}(\varepsilon_r|\varepsilon_{-r}))$$

where $\tilde{\eta}_r = \eta_r - f_r$ when f_r is to be estimated.

Further, $\mathrm{E}(\varepsilon_r|\varepsilon_{-r})$ with $\varepsilon_r = (\varepsilon_{r11}, \ldots, \varepsilon_{rnn_n})^t$ is the conditional mean of the error terms with

$$\mathrm{E}(\varepsilon_{rij}|\varepsilon_{-r,ij}) = \mu_{rij} + \frac{\sigma_{12,ij}}{\sigma_{-r,ij}^2}(y_{-r,ij} - \mu_{-r,ij} - \eta_{-r,ij})$$

and $\Sigma_{r|-r}$ is the conditional covariance matrix with

$$\Sigma_{r|-r} = \mathrm{diag}(\sigma^2_{(r|-r),11}, \ldots, \sigma^2_{(r|-r),nn_n})$$

and

$$\sigma^2_{(r|-r),ij} = \mathrm{Var}(\varepsilon_{rij}|\varepsilon_{-r,ij}) = \sigma^2_{rij} - \frac{\sigma^2_{12,ij}}{\sigma^2_{-r,ij}}.$$

Note that the posterior mean of some function f_r is given by (subject to centering constraints)

$$f_r(\cdot) = (X_r^t \Sigma_{r|-r}^{-1} X_r + \frac{1}{\tau_r^2}\Delta_r)^{-1} X_r^t \Sigma_{r|-r}^{-1}(y_r - \tilde{\eta}_r - \mathrm{E}(\varepsilon_r|\varepsilon_{-r})).$$

It can be noted that the Dirichlet Process Mixture prior induces different variances and therefore $\Sigma_{r|-r}$ weighs observations accordingly just as in the case of heteroscedasticity.

The full conditionals for the smoothing variance parameters τ_r^2, $r = 1,2$ follow inverse Gamma distributions

$$\tau_r^2|\cdot \sim IG(a'_{\tau_r}, b'_{\tau_r})$$

with parameters

$$a'_{\tau_r} = a_{\tau_r} + \frac{\mathrm{rank}(\Delta_r)}{2}, \quad b'_{\tau_r} = b_{\tau_r} + \frac{1}{2}\beta_r^t \Delta_r \beta_r.$$

In the case of the components of the error distribution, the full conditionals can be summarized as the following

- Let $c_i \in \{1,\dots,K^*\}, i=1,\dots,n$ indicate the cluster observation i belongs to.

  For $i=1,\dots,n$:

  - If $c_i = c_h$ for some $h \neq i$, create auxiliary component c^* with (μ_{c^*},Σ_{c^*}) drawn from G_0.
  - If $c_i \neq c_h$ for all $h \neq i$, let $c^* = c_i$ with $(\mu_{c^*},\Sigma_{c^*}) = (\mu_{c_i},\Sigma_{c_i})$.
  - Draw a new value for c_i using

$$\begin{aligned} c_i|c_{-i},y_{1i},y_{2i},\mu_1,\Sigma_1,\dots,\mu_{K^*},\Sigma_{K^*},\mu_{c^*},\Sigma_{c^*} \quad \sim \quad & b\sum_{l=1}^{k^-}\frac{n_l^{-i}}{n-1+\alpha}F\left((\varepsilon_{1i},\varepsilon_{2i}),\mu_l,\Sigma_l\right) \\ & +b\frac{\alpha}{n-1+\alpha}F\left((\varepsilon_{1i},\varepsilon_{2i}),\mu_{c^*},\Sigma_{c^*}\right) \end{aligned}$$

    where k^- is the number of distinct c_h for $h \neq i$, n_l^{-i} is the number of c_h for $h \neq i$ that are equal to l, b is a normalizing constant and $F\left((\varepsilon_{1i},\varepsilon_{2i}),\mu_l,\Sigma_l\right)$ the likelihood for observation i.

- Discard those μ_l,Σ_l that are not associated with one or more observations.

- For all $l \in \{c_1,\dots,c_n\}$: Update μ_l and Σ_l using $\mu_l|\cdot \sim \mathrm{N}(m_{\mu_l},P_{\mu_l}^{-1})$ and $\Sigma_l|\cdot \sim \mathrm{IW}(s'_\Sigma,S'_\Sigma)$ with

$$\begin{aligned} m_{\mu_l} &= (\tau_\Sigma+1)^{-1}\left(\tau_\Sigma\mu_0 + \sum_{i:c_i=l}((y_{1i},y_{2i})-(\eta_{1i},\eta_{2i}))^t\right) \\ P_{\mu_l}^{-1} &= \frac{\tau_\Sigma^{-1}}{1+\tau_\Sigma^{-1}}\Sigma_l/n_l = (\tau_\Sigma+1)^{-1}\Sigma_l/n_l, \\ s'_\Sigma &= s_\Sigma + \frac{n_l}{2} \\ S'_\Sigma &= S_\Sigma + \frac{1}{2}\frac{1}{1+\tau_\Sigma^{-1}}\sum_{i:c_i=l}((y_{1i},y_{2i})-(\eta_{1i},\eta_{2i})-\mu_0)^t\,((y_{1i},y_{2i})-(\eta_{1i},\eta_{2i})-\mu_0) \end{aligned}$$

- The full conditionals of τ_Σ are

$$\tau_\Sigma \sim \mathrm{Ga}\left(\frac{a_\Sigma+K^*}{2},\ \frac{1}{2}\left(b_\Sigma+\sum_{l=1}^{K^*}\Sigma_l^{-1}(\mu_l-\mu_0)^2\right)\right)$$

- The concentration parameter α is drawn from a mixture of two gamma distributions

$$\alpha|\cdot \sim \frac{a_\alpha + K^* - 1}{n(b_\alpha - \log\omega)} \mathrm{Ga}\left(a_\alpha + K^*, b_\alpha - \log\omega\right) + \left(1 - \frac{a_\alpha + K^* - 1}{n(b_\alpha - \log\omega)}\right) \mathrm{Ga}\left(a_\alpha + K^* - 1, b_\alpha - \log\omega\right)$$

where ω is a latent variable sampled from a beta distribution $\omega \sim \mathrm{Be}(\alpha + 1, n)$.

4.4 Performance comparison through simulation

The comparison between the 2SGAM and BNIV approaches, in the context of weak identification scenarios, is accomplished by running Monte Carlo simulation from the same Data Generating Process (DGP) used in Han, 2014.

That DGP is consistent with the model structure described in previous section, and can be sketched as follows

$$y_1 = \pi_0 + \pi_1 w + \varepsilon_1, \quad y_2 = \Phi\left(\frac{y_1 - \mu_{y_1}}{\sigma_{y_1}}\right) + \varepsilon_2,$$
$$\text{where } w \sim N(0,1) \text{ and } (\varepsilon_{1i}, \varepsilon_{2i}) \sim N(0, \Sigma), \tag{4.17}$$
$$\text{with } \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

In DGP (4.17), $\Phi(\cdot)$ is the cumulative function of the standard Normal distribution and variables y_1, y_2 and w are univariate. The degree of endogeneity is controlled by parameter ρ, which we set at 0.5 implying a relatively large correlation between errors. The sample $\{y_{1i}, y_{2i}, w_i\}_{i=1}^n$ is i.i.d with $n = 1000$. We employed a moderated number of $s = 300$ simulation repetitions, due to the high computational burden that BNIV estimations demand.[3]

It is important to note that the joint errors distribution in (4.17) follows a bivariate standard Normal, implying that the true control function, $E(\varepsilon_{2i}|\varepsilon_{1i}) = \rho\varepsilon_{1i}$, is a straight line with slope $\rho = 0.5$ (i.e. the same linear function for all pair of errors). This DGP is a fair alternative to compare 2SGAM versus BNIV, because the true control function is neither a smooth nonlinear

[3] Simulation results obtained using $s = 1000$ for 2SGAM estimators did not differ substantially from the results based on $s = 300$

function (which would favor the former) nor a varying coefficient term (which would favor the later).

The linear specification of the reduced form equation enables us to control the strength of the instrument z through parameter π_1. Moreover, we can use (4.8) to relate the concentration parameter with coefficient π_1 as follows

$$\gamma^2 = \pi_1^2 \sum_{i=1}^{n} w_i^2,$$

which presents an expected value equal to $\gamma^2 = \pi_1^2 n$.

We establish the sequence of values $\{4, 10, 12, 16, 32, 64, 256\}$ for the concentration parameter, that includes the parametric threshold $\gamma^2 = 10$ and values ranging from weak to strong instruments. Finally, we set the reduced form intercept π_0 equal to 2.

Even though the reduced form equation can be estimated using a linear model, and the control function can be modeled with a linear term, we keep the skepticism about any particular functional form, and estimate them in a flexible way (as is the case in Han, 2014).

Table 4.1 presents integrated squared bias, integrated variance and integrated mean squared error (MSE) of the 2SGAM estimator (for both GCV and REML methods of smoothing parameter selection) and the GAM estimator using GCV method which constitutes the *naive* estimator (i.e. without endogeneity correction). Additionally, ratios of integrated MSE are reported for comparisons. All estimators was performed using a number of 20 knots for the spline basis functions. Additional sensitivity analysis showed that using alternative basis functions (i.e. modifying the number of knots and/or using different types of spline basis) does not significantly change the estimation results.

For decreasing degrees of IV's strength until $\gamma^2 = 10$, the computed squared bias maintains relatively low and stable values for both estimators $2SGAM_{GCV}$ and $2SGAM_{REML}$, but variance increases in an accelerated way. In general, it can be noted that $2SGAM_{GCV}$ presents lower bias and higher variance than $2SGAM_{REML}$, but differences are small for values of 10, 12 and 16 of the concentration parameter γ^2.

In terms of integrated MSE, and for γ^2 smaller than or equal to 10, both 2SGAM estimators perform worse than the naive GCV ($Naive_{GCV}$) estimator. This reveals the necessity to establish a higher value for γ^2 as the new threshold for characterizing nonparametric weak identification scenarios, instead to the threshold of $\gamma^2 = 10$ established in the parametric case. For γ^2 equal to 12, 16 and 32, integrated MSE of 2SGAM estimators represents about 60%, 40% and 15% of the $Naive_{GCV}$'s MSE, respectively. These results suggest that the new threshold for the concentration parameter may be specified as a value between 12 and 16.

TABLE 4.1: *Integrated squared bias, integrated variance and integrated mean squared error of the naive and 2SGAM estimators for the term* $\Phi(\cdot)$.

					γ^2			
Estimator		4	10	12	16	32	64	256
	Bias^2	0.649	0.013	0.018	0.025	0.027	0.024	0.018
$2SGAM_{GCV}$	*Var*	60.93	6.129	3.114	1.779	0.742	0.385	0.148
	MSE	61.58	6.142	3.132	1.803	0.769	0.409	0.158
	Bias^2	0.571	0.022	0.024	0.029	0.035	0.033	0.028
$2SGAM_{REML}$	*Var*	61.83	6.059	3.001	1.662	0.661	0.335	0.129
	MSE	62.40	6.082	3.025	1.691	0.696	0.369	0.166
	Bias^2	4.896	4.869	4.860	4.853	4.7840	4.6297	3.9672
$Naive_{GCV}$	*Var*	0.083	0.081	0.078	0.081	0.080	0.085	0.080
	MSE	4.979	4.949	4.940	4.934	4.864	4.714	4.047
$\frac{2SGAM_{GCV}MSE}{Naive_{GCV}MSE}$		12.37	1.24	0.63	0.37	0.16	0.09	0.04
$\frac{2SGAM_{REML}MSE}{Naive_{GCV}MSE}$		12.53	1.23	0.61	0.38	0.14	0.08	0.04

Simulation results in Table 4.1 differ from findings in Han, 2014 which shows that the IV estimator behaves worse than the naive one even for $\gamma^2 = 16$. This fact can be explained by the lack of a data driven method for optimal smoothing parameter selection in Han, 2014's IV estimators.

Table 4.2 reports the same performance scores of Table 4.1 but for the BNIV estimator. Compared to the 2SGAM and for γ^2 ranging from 10 to 16, the BNIV shows equal or smaller squared bias and a smaller variance with a notable smoother behavior. This enables the BNIV estimator to present a relative MSE (over the $Naive_{GCV}$' MSE) of 45%, 40% and 30% for γ^2 equal to 10, 12 and 16 respectively. This result suggests that the parametric weak identification threshold of $\gamma^2 = 10$ can be valid in the case of BNIV.

Table 4.3 presents relative comparisons between BNIV and 2SGAM estimators, in terms of bias, variance and mean squared error. For $\gamma^2 \leq 16$ BNIV outperforms the 2SGAM estimators (for both GCV and REML versions) in terms of bias and variance. This situation reverts when

TABLE 4.2: *Integrated squared bias, integrated variance and integrated mean squared error of the BNIV estimator of the term $\Phi(\cdot)$.*

		γ^2						
Estimator		4	10	12	16	32	64	256
BNIV	*Bias*^{2}	0.409	0.013	0.006	0.002	0.002	0.003	0.007
	Var	4.292	2.212	1.958	1.464	0.793	0.440	0.191
	MSE	4.701	2.225	1.963	1.466	0.795	0.443	0.198
$\frac{BNIV MSE}{Naive_{GCV} MSE}$		0.94	0.45	0.40	0.30	0.16	0.09	0.05

concentration parameter takes values of 32 or larger.

TABLE 4.3: *Performance comparison between BNIV and 2SGAM estimators of the term $\Phi(\cdot)$.*

	γ^2						
Estimator	4	10	12	16	32	64	256
$\frac{BNIV Bias^2}{2SGAM_{REML} Bias^2}$	0.72	0.59	0.25	0.07	0.06	0.09	0.25
$\frac{BNIV Var}{2SGAM_{REML} Var}$	0.07	0.36	0.65	0.88	1.19	1.31	1.48
$\frac{BNIV MSE}{2SGAM_{REML} MSE}$	0.08	0.36	0.65	0.87	1.14	1.20	1.19
$\frac{BNIV Bias^2}{2SGAM_{GCV} Bias^2}$	0.63	1	0.33	0.08	0.07	0.125	0.39
$\frac{BNIV Var}{2SGAM_{GCV} Var}$	0.07	0.36	0.63	0.82	1.07	1.14	1.29
$\frac{BNIV MSE}{2SGAM_{GCV} MSE}$	0.08	0.36	0.63	0.81	1.03	1.08	1.25

Previous considerations support the idea that BNIV must be the preferred option when potentially weak instruments are present (i.e. when $10 \leq \gamma^2 \leq 16$), mainly because of its ability to mitigate the estimations variability induced by the concurvity problem.

To get a visual comparison between BNIV and 2SGAM, Figure 4.1 presents the mean and 0.03-0.97 quantile range for estimations of the function $\Phi(\cdot)$ with concentration parameter equal to 10, 12, 16 and 32.

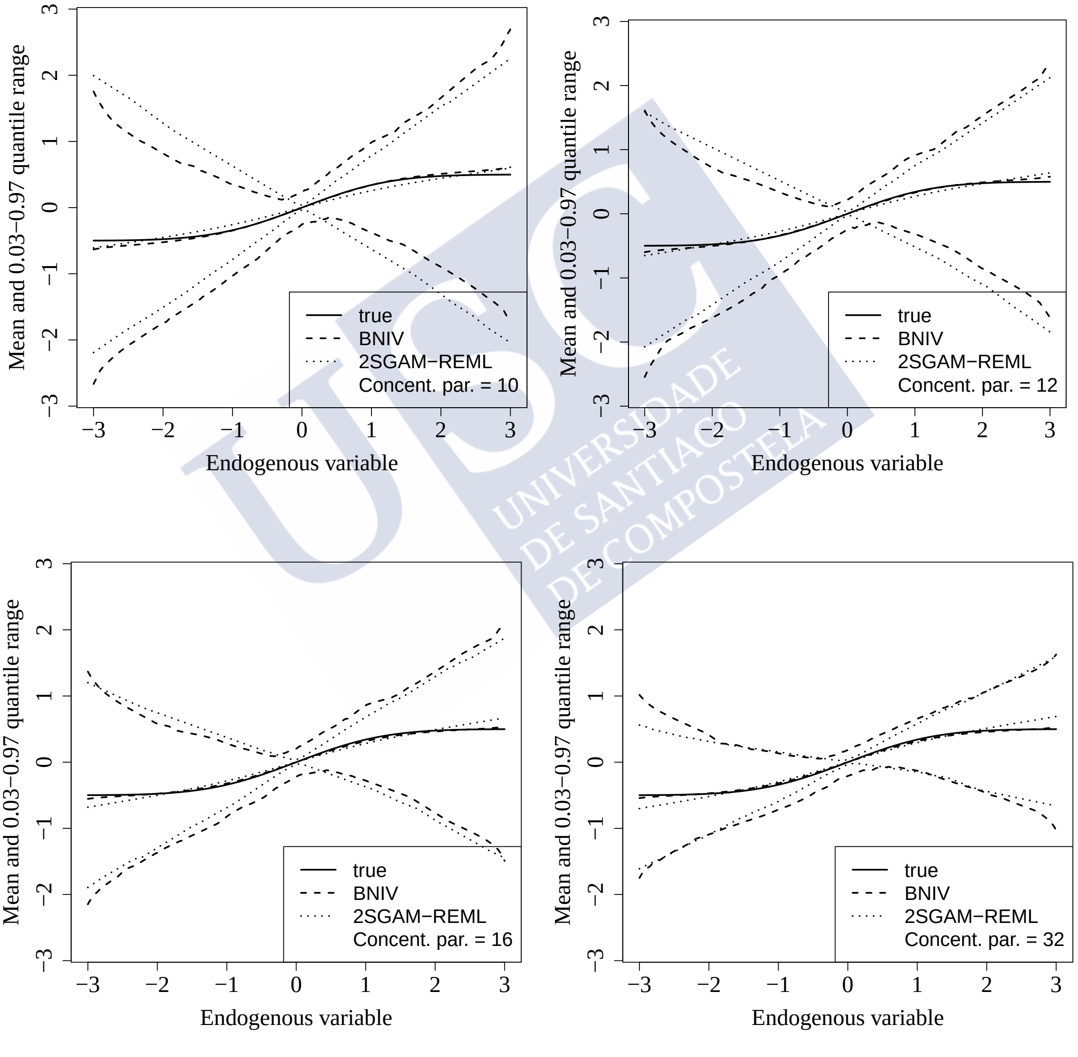

FIGURE 4.1: *Mean and 0.03-0.97 quantile range of BNIV (dashed line) and 2SGAM (dotted line) estimations of $\Phi(\cdot)$ with γ^2 equal to 10 (top-left plot), 12 (top-right plot), 16 (bottom-left plot) and 32 (bottom-right plot).*

When $\gamma^2 = 10, 12$, the variance of 2SGAM is excessively large, producing a quantile range which may include a function of any shape. On the other hand, the variance of BNIV is significantly smaller implying a quantile range that preserves the positive shape of the true function $\Phi(\cdot)$.

In the case of $\gamma^2 = 16$, 2SGAM improves its variance performance, but BNIV still remains superior in terms of variance.

It is remarkable that the above-mentioned higher efficiency of BNIV does not come at the cost of a larger bias, in fact the bias of BNIV maintains stable values which are in general smaller than the 2SGAM's bias. Finally, when the concentration parameter is equal to 32, BNIV and 2SGAM estimators perform in a similar way in terms of variance and quantile ranges.

A criticism that can be made to the previous comparison using DGP (4.17), is that a joint normal distribution is assumed for error terms $(\varepsilon_{1i}, \varepsilon_{2i})$, which can favour estimation trough BNIV.

To ensure a fairer comparison between BNIV and 2SGAM, a slight modification to the DGP (4.17) is introduced. In DGP (4.18) joint normality of errors is replaced by a Uniform distribution for ε_{1i} and a mixture of Uniform and Normal distributions for ε_{2i}.

$$\begin{gathered} y_1 = \pi_0 + \pi_1 w + \varepsilon_1, \quad y_2 = \Phi\left(\frac{y_1 - \mu_{y_1}}{\sigma_{y_1}}\right) + \varepsilon_2, \\ \text{where } w \sim N(0,1)\,, \varepsilon_{1i} \sim Unif(-1,7,1,7)\,, \varepsilon_{2i} = 0.5\varepsilon_{1i} + \varepsilon_{3i}, \\ \text{with } \varepsilon_{3i} \sim N(0, 0.757). \end{gathered} \tag{4.18}$$

Table 4.4 present the comparison between 2SGAM and BNIV estimators in the context of DGP (4.18) and simulated samples $\{y_{1i}, y_{2i}, w_i\}_{i=1}^n$ with $n = 200$. In this case we increased the number of simulation repetitions to $s = 600$.

Simulation results confirm the previous conclusions about the advantages of BNIV in terms of lower bias and efficiency gains when compared to 2SGAM (for both versions GCV and REML). In addition, new results extends such BNIV superiority to scenarios when the concentration parameter values are larger than 16 (i.e. $\gamma^2 => 16$). For example, with $\gamma^2 = 64$ 2SGAM-REML is 58% larger than BNIV in terms of integrated mean squared error.

TABLE 4.4: *Integrated squared bias, integrated variance and integrated mean squared error of 2SGAM and BNIV estimators for the term $\Phi(\cdot)$ in DGP (4.18).*

		γ^2				
Estimator		10	16	32	64	128
	$Bias^2$	4.683	4.452	3.875	3.262	2.670
$Naive_{GCV}$	Var	1.673	1.411	0.895	0.773	0.690
	MSE	6.355	5.863	4.770	4.035	3.360
	$Bias^2$	0.079	0.079	0.071	0.062	0.059
$2SGAM_{GCV}$	Var	5.602	3.062	1.396	1.009	0.643
	MSE	5.681	3.141	1.467	1.071	0.702
	$Bias^2$	0.046	0.044	0.062	0.070	0.073
$2SGAM_{REML}$	Var	4.267	2.048	0.832	0.479	0.327
	MSE	4.314	2.093	0.894	0.549	0.400
	$Bias^2$	0.026	0.028	0.026	0.019	0.015
$BNIV$	Var	0.812	0.629	0.441	0.328	0.277
	MSE	0.838	0.658	0.467	0.347	0.292
$\frac{2SGAM_{GCV}MSE}{BNIV MSE}$		6.78	4.77	3.14	3.08	2.41
$\frac{2SGAM_{REML}MSE}{BNIV MSE}$		5.15	3.18	1.91	1.58	1.37

4.5 Results discussion

This chapter assesses the performance of two alternative flexible estimators of the Triangular Simultaneous Equations Model when weak instruments are present. Both analyzed estimators, the Two Stage Generalized Additive Model (2SGAM) and the Bayesian Nonparametric Instrumental Variables (BNIV), are of greater relevance for applied research because they enable data-driven smoothing parameter selection and provide valid tools to perform statistical inference.

Simulation results support the advantages of BNIV over 2SGAM when instruments are weak. Specifically, when the concentration parameter ranges between 10 and 16, BNIV outperform 2SGAM in terms of variance. This finding is not surprising because it is expected that the flexible structure of the control function term, implied by BNIV, helps to reduce the concurvity problem associated with weak identification.

It is important to note that the mentioned efficiency of BNIV does not imply an increment in relative bias. In fact, bias in BNIV remains significantly smaller than 2SGAM's bias when concentration parameter is larger than 10.

The issue of the minimum value of the concentration parameter required to avoid the nonparametric weak identification problem, in the context of additive frequentist estimators as 2SGAM, needs to be further studied. The suggestion made in Han, 2014 about that the concentration parameter must exceed the value of 16 is called into question based on our simulation results. We find that in the 12-16 range of the concentration parameter, the 2SGAM estimator presents an acceptable performance even though it is outperformed by BNIV estimator.

Focusing on the BNIV estimator, it seems to yield useful estimation results in terms of identification of the unknown function to be estimated, even when the concentration parameter is equal to 10.

Comparing alternative approaches of smoothing parameter selection in the 2SGAM case, we find that REML method is more efficient than GCV method, but the later presents lower bias.

Beyond the results registered in this chapter characterizing the weak identification problem, further research must be done to assess the performance of uncertainty measures, as confidence/credible interval construction, for both methodologies, BNIV and 2SGAM .

Appendix A

Chapters appendices

This final chapter has two goals. First, it introduces some additional explanations and estimation details which complete the analysis of previous chapters. Second, it presents a part of the R (R Core Team, 2014) code used in each of the previous chapters. This selected code might be useful for practitioners trying to replicate some of the addressed analysis or for future construction of functions that automate the estimation and inference procedures, originating a new R package.

A.1 Appendix of Chapter 2

A.1.1 Estimation details

Estimation of Additive Models (AM) and Generalized Additive Models (GAM) involves several aspects that need to be carefully considered by the researcher. Particularly, the methodology to select or estimate the so called *smoothing parameters* is of primary concern.

The flexible models (AMs and GAMs) in Chapter 1 have been estimated by Restricted Maximum Likelihood (REML) method, using the 'gam()' function of the R package 'mgcv' (Wood, 2011, Wood, 2006a and Wood, 2004), which estimates the smoothing parameters from the data. This method has been chosen over several alternatives, including Generalized Cross Validation (GCV), Akaike Information Criterion (AIC) or Un-Biased Risk Estimator (UBRE), due to specific advantages like the improvement in terms of the mean square error and in stability in the presence of severe under-smoothing failures (Wood, 2011 and Wood, 2013). In fact, the estimation results using default method GCV (Generalized Cross Validation) in the present application tend to show an under-smoothing behavior for some model components, including *TDC* effect. All the others options in the 'gam()' function have been set to the default values.

The construction of point-wise confidence intervals is based on the Bayesian view of the smoothing process, as proposed in Wood, 2006b and studied by Marra and Wood, 2012. Such Bayesian approach enables us to test the significance of smooth terms through the computation of Wald type statistics with their corresponding p-values (Wood, 2013). We followed Marra and Radice, 2011 which extends this Bayesian methodology to be used in the simultaneous equations framework.

Another problem we dealt with was the heteroscedastic error's structure, present in the structural equation 2.10. To deal with that problem we extended the procedure known as Feasible Generalized Least Squared, commonly applied to linear models (Wooldridge, 2010), to the Additive Model context. In concrete, we exploit the fact that variance of the binary outcome variable *Event* is given by

$$P(Event = 1|TDC, ..., u)P(Event = 0|TDC, ..., u),$$

which can be consistently estimated by 2SAM or 2SGAM.

Then, to correctly estimate the Variance-Covariance matrix of estimated coefficients, we have used the 2SAM procedure but applying the following weights in both stages

$$\frac{1}{\sqrt{\hat{P}(Event = 1|TDC, ...)\hat{P}(Event = 0|TDC, ...)}},$$

where the predicted probability $\hat{P}(Event = 1|...)$ was obtained by 2SGAM (with a Probit link function) to assure non-negative predicted values. Then, the resulting Variance-Covariance matrix was used to compute the Bayesian Confidence Intervals.

A similar weighted procedure has been applied to estimate Bayesian (heteroscedasticity robust) standard errors in the linear model case, fitted through 2SLS. The estimates obtained are very similar to the robust standard errors produced by Hubert-White asymptotic formula reported in Table 2.7.

A.1.2 The R environment routines

In this subsection we introduce part of the R code used in Chapter 2. Estimation results for flexible models with endogeneity correction can be obtained from this code. The results related to the linear model and the GLM cases, can be obtained by R packages 'AER' and 'ivpack', or by Stata (Statistics/Data Analysis - StataCorp) command 'ivregress' and 'ivprobit'. The software versions used in the analysis are R 3.0.3 and Stata 12.

The variables described in Chapter 2 possess a different name in the dataset, the new names are TDC = horas-delay, Age = EDAD, Gr = GRS-iH, Fem = sex, Fr = IVvie, Sa = IVsab and Su = IVdom.

```
### Data ###
```

```
 library(foreign)
 Data<-read.dta("C:/.../Cate2014_conIVdomOK_Completa.dta")
 Dat1<-Data[Data$horas_delay<=60,]

 library(mgcv)
 library(mvtnorm)

####  2.5 Empirical Results ####
################################

## Table 2.9 and Figure 2.2: First stage
 fs<-gam(horas_delay~IVsab+IVdom+IVvie+s(GRS_iH)+s(EDAD)
  +sex, method="REML", data=Dat1, na.action="na.exclude")
 summary(fs)
 Dat1$resi <- residuals(fs, type = c("response"))

## Table 2.10, part I:
#  Second stage without heteroscedasticity correction
#(to get consistent coefficients estimation)
 ss<-gam(Ex_IAM~s(horas_delay)+s(EDAD)+s(GRS_iH)+sex
  +s(resi), method="REML", data=Dat1)
 summary(ss)

## Table 2.10, part II:

#  Second stage with heteroscedasticity correcton
# (to get robust confidence intervals)
 ssPr<-gam(Ex_IAM~s(horas_delay)+s(EDAD)+s(GRS_iH)+
 sex+s(resi), method="REML", data=Dat1,
 family=binomial(link="probit"))
 Pr.hat<-predict(ssPr, type="response")
 Dat1$weig <- 1 / (sqrt(Pr.hat*(1-Pr.hat)))

 ## First stage using weights
 fsh<-gam(horas_delay~IVsab+IVdom+IVvie+s(GRS_iH)+
 s(EDAD)+sex, method="REML", data=Dat1, weights=weig,
 na.action="na.exclude")
 Dat1$resih <- residuals(fsh, type = c("response"))
 ## Second stage using weights
 ssh<-gam(Ex_IAM~s(horas_delay)+s(EDAD)+s(GRS_iH)+sex
  +s(resih) , method="REML", data=Dat1, weights=weig)
 summary(ssh)


###### Bayesian Confidence Bands ######

yf <- fs$y
n.boot <- 100
n.draw <- 400
XE <- model.matrix(fs)
beta <- matrix(NA, length(coef(ss)), n.boot*n.draw)

coe.list <- list()
var.list <- list()
coe.list[[1]] <- coef(ss)
var.list[[1]] <- ssh$Vp

for(k in 1:(n.boot-1)){
  xe.star <- XE%*%t(rmvnorm(1 , fs$coeff, fs$Vp ))
  res <- yf - xe.star
  xe.star2 <- XE%*%t(rmvnorm(1 , fs$coeff, fsh$Vp ))
  res2 <- yf - xe.star
  # Second Stage without heteroscedasticity correction
  #(to get simulated coefficients)
  sstsp<-gam(Ex_IAM~s(horas_delay)+s(EDAD)+s(GRS_iH)+sex
  +s(res), method="REML", data=Dat1, na.action="na.exclude")
  # Second Stage without heteroscedasticity correction
  # (to get Robust Var-Covar Matrix)
  sst <-gam(Ex_IAM~s(horas_delay)+s(EDAD)+s(GRS_iH)+sex+
   s(res2), method="REML",data=Dat1,weights=weig,
   na.action="na.exclude")
  coe.list[[k+1]] <- coef(sstsp)
  var.list[[k+1]] <- sst$Vp
  print(k)
}

for (k in 1:n.boot){
 beta[, ((k-1)*n.draw+1):(k*n.draw) ] <- t(rmvnorm(n.draw,
    coe.list[[k]], var.list[[k]] ))
}

LB1 <- UB1 <- LB2 <- UB2 <- LB3 <- UB3 <- LB4 <- UB4 <- NA
```

```
sm <- 1
ind<-ss$smooth[[sm]]$first.para:ss$smooth[[sm]]$last.para
Xq <- model.matrix(ss)[,ind]
Bq <- beta[ind,]
F.q <- Xq%*%Bq

for(i in 1:dim(F.q)[1]){
LB1[i] <- quantile(F.q[i,],c(.025))
UB1[i] <- quantile(F.q[i,],c(.975))
}

sm <- 2
ind2<-ss$smooth[[sm]]$first.para:ss$smooth[[sm]]$last.para
Xq2 <- model.matrix(ss)[,ind2]
Bq2 <- beta[ind2,]
F.q2 <- Xq2%*%Bq2

for(i in 1:dim(F.q)[1]){
LB2[i] <- quantile(F.q2[i,],c(.025))
UB2[i] <- quantile(F.q2[i,],c(.975))
}

sm <- 3
ind3<-ss$smooth[[sm]]$first.para:ss$smooth[[sm]]$last.para
Xq3 <- model.matrix(ss)[,ind3]
Bq3 <- beta[ind3,]
F.q3 <- Xq3%*%Bq3

for(i in 1:dim(F.q)[1]){
LB3[i] <- quantile(F.q3[i,],c(.025))
UB3[i] <- quantile(F.q3[i,],c(.975))
}

sm <- 4
ind4<-ss$smooth[[sm]]$first.para:ss$smooth[[sm]]$last.para
Xq4 <- model.matrix(ss)[,ind4]
Bq4 <- beta[ind4,]
F.q4 <- Xq4%*%Bq4

for(i in 1:dim(F.q)[1]){
```

```
LB4[i] <- quantile(F.q4[i,],c(.025))
UB4[i] <- quantile(F.q4[i,],c(.975))
}


### Robust standard errors and confidence intervals
#   of parametric components:
LBpa <- UBpa <- NA
beta.pa <- beta[1:2,]

for(i in 1:2){
LBpa[i] <- quantile(beta.pa[i,],c(.025))
UBpa[i] <- quantile(beta.pa[i,],c(.975))
}

# CIs
round(cbind(t(t(LBpa)),summary(ss)$p.coeff,t(t(UBpa))),4)

# Std Error
sqrt(var(beta.pa[1,]))  # intercept
sqrt(var(beta.pa[2,]))  # sex


### First derivative of TDC: finite difference approx. ###
##########################################################

# new data for prediction
 newDat<-with(Dat1, data.frame(horas_delay=
 unique(horas_delay)))
 ng<-length(newDat[,1])
 newDat$EDAD<-seq(min(Dat1$EDAD), max(Dat1$EDAD),
  length=ng)
 newDat$GRS_iH<-seq(min(Dat1$GRS_iH),max(Dat1$GRS_iH),
  length=ng)
 newDat$resi<-seq(min(Dat1$resi), max(Dat1$resi),
  length=ng)
 newDat$sex<-rep(0,ng)  ##  male patients only

# finite difference
eps <- 1e-07
X0 <- predict(ss, newDat, type = ’lpmatrix’)
```

```
newDat_p <- newDat + eps
newDat_p$sex<-0
X1 <- predict(ss, newDat_p, type = ’lpmatrix’)

# finite difference approximation of first derivative
# the design matrix
xp <- (X1 - X0) / eps

# first derivative for treatment TDC
sm <- 1
ind<-ss$smooth[[sm]]$first.para:ss$smooth[[sm]]$last.para
xq <- xp[,ind]
bq <- coef(ss)[ind]
d1_F.q <- xq%*%bq

### First derivative bayesian C.I.

Bq <- beta[ind,]
D1_F.q <- xq%*%Bq

dLB1 <- dUB1 <- dLB2 <- dUB2 <- dLB3 <- dUB3 <- NA

for(i in
1:dim(D1_F.q)[1]){
dLB1[i] <- quantile(D1_F.q[i,],c(.025))
dUB1[i] <- quantile(D1_F.q[i,],c(.975))
#dLB199[i] <- quantile(D1_F.q[i,],c(.01))
#dUB199[i] <- quantile(D1_F.q[i,],c(.99))
}

## Figure 2.1 (right)
windows()
plot(sort(newDat$horas_delay),
  d1_F.q[order(newDat$horas_delay)], type="lines",
  cex.axis=1.0, cex.main=1.3, ylim=c(-0.006,0.02),
  col="grey60" , lwd=1.7, cex.lab=1.6, cex.axis=1.6,
  ylab="f ’ (TDC)", xlab="TDC (hours)")
 lines(sort(newDat$horas_delay),
  dLB1[order(newDat$horas_delay)], col="grey60",
  lty="dashed", lwd=1.7)
 lines(sort(newDat$horas_delay),
  dUB1[order(newDat$horas_delay)], col="grey60",
  lty="dashed", lwd=1.7)
 abline(h=0, lty="dotted")



### TDC Average First Derivative (Marginal Effect: MgEf) ###
#############################################################

## Marginal effect for the whole TDC range
xx <- model.matrix(ss)
eps <- 1e-07
XX0 <- predict(ss, Dat1, type = 'lpmatrix')
dat1<- with(Dat1, data.frame(EDAD, horas_delay, GRS_iH,
 Ex_IAM,resi, sex))
dat1_p <- dat1 + eps
XX1 <- predict(ss, dat1_p, type = 'lpmatrix')
xxp <- (XX1 - XX0) / eps

sm <- 1
ind<-ss$smooth[[sm]]$first.para:ss$smooth[[sm]]$last.para
xx1 <- xxp[,ind]
bb1 <- coef(ss)[ind]
d1_ff1 <- xx1%*%bb1

mean(d1_ff1)

## Marginal effect for TDC between 0 and 34
dat2<- with(Dat1, data.frame(EDAD, horas_delay, GRS_iH,
 Ex_IAM, resi, sex))[Dat1$horas_delay<=34,]
XX02 <- predict(ss, dat2, type = 'lpmatrix')
dat2_p <- dat2 + eps
XX2 <- predict(ss, dat2_p, type = 'lpmatrix')
xxp2 <- (XX2 - XX02) / eps

sm <- 1
ind<-ss$smooth[[sm]]$first.para:ss$smooth[[sm]]$last.para
xx12 <- xxp2[,ind]
bb1 <- coef(ss)[ind]
d1_ff2 <- xx12%*%bb1
```

```
mean(d1_ff2)
```

A.2 Appendix of Chapter 3

A.2.1 Estimation methodology for flexible models

All flexible models in Chapter 3, single Additive Models (AM) and Two Stage Additive Models (2SAM), have been estimated by Restricted Maximum Likelihood (REML) method, using ‘gam()’ function of R package ‘mgcv’ (Wood, 2011), which estimates the smoothing parameters from the data. As we mentioned in subsection 5.1.1, the REML estimator is sometimes preferred over Generalized Cross Validation (GCV), Akaike Information Criterion (AIC) or Un-Biased Risk Estimator (UBRE) because it presents advantages like some improvement in the mean squared error and robustness to occasional severe under-smoothing (Wood, 2011 and Wood, 2013).

We restricted the Empirical Degrees of Freedom (e.d.f.) values, for all smooth effects terms in all the estimated models, to a maximum of 5. This restriction was necessary to prevent overfitting. Such restriction is based on theory grounds, which expects the existence of smooth effects of covariates.

The degrees of freedom restriction was more important for covariate Enrollment, which was limited to a maximum of 2 e.d.f., because instrument *PCS* is in fact a deterministic (highly) nonlinear function of Enrollment.

To run the outlier detection strategies, based on Stahel, 1981 and Donoho, 1982 univariate projections estimator, we rely on the function "outlyingness" in R package "mrfDepth". That function presents alternatives options to compute the robust standardised distance. We choose default option with which the data for each projection is centred around the median and standardised by the median absolute deviation.

A.2.2 The R routine for analysis reproducibility

In this subsection we present the R (R Core Team, 2014) code used in Chapter 3. Regression results, including bootstrap inference, similar to reported in Chapter 3, can be obtained from this routine. The software version used in the analysis was R 3.0.3. The statistical program Stata 12 was used to compute asymptotic robust standard errors for the estimated coefficients in the parametric models (routines not reported).

The variables referred to in Chapter 3 possess a different name in the dataset, new names are *Score Literature* = score100-len, *Class Size* = numtest-ok, *Predicted Class Size* = IVxx, *Enrollment* = smaxalu1, *Socioeconomic Index* = sxf31sal, *High education* = por-edu-alta, *Housing Issues* = por-prob-espac and *Repeaters* = por-repit.

```
######  Data reading  ######
############################

library(mgcv)

library(foreign)
Unif.len<-read.dta("Unif96_final.dta", convert.factors = TRUE)
Unif.len<-as.data.frame(Unif.len)
#attach(Unif.len)

#### Sample excluding Outlier Schools

# Outliers exclusion from data
X<-matrix(c(Unif.len$numtest_ok, Unif.len$IVxx),nrow=4744, ncol=2)

library(mrfDepth)
OutAn<-outlyingness(x=X)

 Unif.len2<-Unif.len[OutAn$flagX,]
 attach(Unif.len2)


### Clusters ID (used hereafter for all regressions based on
#     full sample)
ruees1<-c(1101001,1101002,1101003,1101005,1101006,1101017,1101018,
1101024,1101025,1101027,1101028,1101029,1101031,1101036,
1101037,1101039,1101040,1101046,1101048,1101049,1101050,
1101052,1101054,1101055,1101061,1101062,1101066,1101069,
1101073,1101074,1101075,1101076,1101077,1101079,1101081,
1101082,1101083,1101084,1101085,1101086,1101087,1101090,
1101092,1101094,1101098,1101100,1101101,1101103,1101104,
1101105,1101107,1101108,1101110,1101112,1101114,1101115,
1101116,1101117,1101118,1101120,1101122,1101124,1101127,
1101128,1101131,1101132,1101137,1101141,1101146,1101147,
1101149,1101150,1101154,1101155,1101156,1101164,1101166,
1101167,1101170,1101171,1101172,1101173,1101174,1101175,
1101180,1101181,1101190,1101192,1101195,1101249,1101251,
1101255,1101258,1101262,1101263,1101266,1101267,1101268,
1101270,1101274,1101277,1101283,1101290,1101299,1101302,
1101309,1101317,1101321,1101323,1101330,1101339,1101344)


### Clusters ID (used hereafter for all regressions based on
#     trimmed sample)
ruees <- c(1101001,1101002,1101003,1101005,1101006,1101017,
1101018,1101024,1101025,1101027,1101028,1101029,1101031,
1101036,1101037,1101039,1101040,1101046,1101048,1101049,
1101050,1101052,1101055,1101061,1101062,1101066,1101069,
1101073,1101074,1101075,1101076,1101077,1101079,1101081,
1101082,1101083,1101086,1101087,1101090,1101092,1101094,
1101098,1101100,1101101,1101103,1101105,1101107,1101108,
1101110,1101112,1101114,1101115,1101116,1101117,1101118,
1101120,1101124,1101127,1101131,1101132,1101137,1101141,
1101146,1101147,1101150,1101154,1101155,1101156,1101164,
1101166,1101167,1101170,1101171,1101172,1101173,1101174,
1101175,1101181,1101190,1101192,1101249,1101255,1101258,
1101262,1101263,1101266,1101270,1101283,1101290,1101299,
1101302,1101309,1101321,1101323,1101330,1101339,1101344)



####    2SAM  ->  endogeneity correction    ####
################################################

#### 1. 2SAM with full sample

kx <- 6  # degrees of freedom restriction

# First stage (reduced form equation)
fs.len<-gam(numtest_ok~ s(smaxalu1, k=3)+s(sxf31sal, k=kx)
+s(por_repit, k=kx)+s(por_edu_alta,k=kx)+
s(por_prob_espac,k=kx)+ IVxx, method = "REML",
data=Unif.len)
 resid.fs.len<-fs.len$residuals
 summary(fs.len)
```

```
# Second stage (reduced form equation)
ss.len<-gam(score100_len~ s(smaxalu1, k=3)+  numtest_ok
+s(sxf31sal, k=kx)+s(por_repit, k=kx)+s(por_edu_alta,k=kx)
+s(por_prob_espac, k=kx) +  s(resid.fs.len), method="REML",
 data=Unif.len)
 summary(ss.len)


####  Weighted bootstrap   ####

B <- 1000
N1 <- length(Unif.len$ruee)

nescu1 <- length(ruees1)
weights1<-c()
Unif.len$weights1<-NaN

cseff1<-c()
interceptSE <- c()
interceptPE <- c()
IV <- c()

 for (i in 1:B)  {

 set.seed(i)
  weights1 <- sqrt(-log(1-runif(nescu1)) )

for (j in 1:N1) {

for (k in 1:nescu1) {
if (Unif.len$ruee[j]==ruees1[k])
Unif.len$weights1[j]<-weights1[k]
  }
}

 fs<-gam(numtest_ok~ s(smaxalu1, k=3)+ s(sxf31sal, k=kx)
     + s(por_repit, k=kx)+ s(por_edu_alta, k=kx)
     + s(por_prob_espac, k=kx) + IVxx , method ="REML",
         data=Unif.len, weights=Unif.len$weights1)
  resid.fs<-fs$residuals
 ss<-gam(score100_len~ s(smaxalu1, k=3)+  numtest_ok
 + s(sxf31sal, k=kx)+s(por_repit, k=kx)+s(por_edu_alta,k=kx)
 + s(por_prob_espac,  k=kx) + s(resid.fs), method ="REML",
 data=Unif.len, weights=Unif.len$weights1)

  cseff1[i] <- ss$coeff[[2]]
  interceptSE[i] <- ss$coeff[[1]]

  interceptPE[i] <- fs$coeff[[1]]
  IV[i] <- fs$coeff[[2]]

 }


### Second stage Confidence Intervals and standard errors

# CS effect

# 1%
q005<-quantile(cseff1,0.005, na.rm=T)
q995<-quantile(cseff1,0.995, na.rm=T)
q005;q995
# 5%
q025<-quantile(cseff1,0.025, na.rm=T)
q975<-quantile(cseff1,0.975, na.rm=T)
q025;q975

ee.cseff1<- sqrt(var(cseff1))
ee.cseff1

# Intercept

# 1%
q005<-quantile(interceptSE,0.005, na.rm=T)
q995<-quantile(interceptSE,0.995, na.rm=T)
q005;q995
# 5%
q025<-quantile(interceptSE,0.025, na.rm=T)
q975<-quantile(interceptSE,0.975, na.rm=T)
q025;q975
```

```
ee.intSE <- sqrt(var(interceptSE))
ee.intSE


### Second stage Confidence Intervals and standard errors

# IV (PCS)

# 1%
q005<-quantile(IV,0.005, na.rm=T)
q995<-quantile(IV,0.995, na.rm=T)
q005;q995
# 5%
q025<-quantile(IV,0.025, na.rm=T)
q975<-quantile(IV,0.975, na.rm=T)
q025;q975

ee.IV <- sqrt(var(IV))
ee.IV

# Intercept

# 1%
q005<-quantile(interceptPE,0.005, na.rm=T)
q995<-quantile(interceptPE,0.995, na.rm=T)
q005;q995
# 5%
q025<-quantile(interceptPE,0.025, na.rm=T)
q975<-quantile(interceptPE,0.975, na.rm=T)
q025;q975

ee.intPE <- sqrt(var(interceptPE))
ee.intPE



#### 2. 2SAM excluding  outliers

# First stage (reduced form equation)
fs.len2<-gam(numtest_ok~ s(smaxalu1, k=3)+s(sxf31sal,k=kx)
  + s(por_repit, k=kx)+ s(por_edu_alta, k=kx)
  + s(por_prob_espac, k=kx) + IVxx , method = "REML",
  data=Unif.len2)
 resid.fs.len2<-fs.len2$residuals
 summary(fs.len2)

 ss.len2<-gam(score100_len~ s(smaxalu1, k=3)+ numtest_ok
 +s(sxf31sal, k=kx)+s(por_repit, k=kx)+s(por_edu_alta,k=kx)
 +s(por_prob_espac, k=kx)+s(resid.fs.len2), method="REML",
 data=Unif.len2)
 summary(ss.len2)


####  Weighted Bootstrap   ####

B<-1000
N<-length(Unif.len2$ruee)

nescu <- length(ruees)
weights<-c()
Unif.len2$weights<-NaN

cseff<-c()
interceptSE2 <- c()
interceptPE2 <- c()
IV2 <- c()



 for (i in 1:B)  {

 set.seed(i)
  weights <- sqrt(-log(1-runif(nescu)) )

for (j in 1:N) {

for (k in 1:nescu) {
if (Unif.len2$ruee[j]==ruees[k])
 Unif.len2$weights[j]<-weights[k]
  }
}
```

```
 fs<-gam(numtest_ok~ s(smaxalu1, k=3)+s(sxf31sal, k=kx)
     + s(por_repit, k=kx)+ s(por_edu_alta, k=kx)
     + s(por_prob_espac, k=kx) + IVxx , method = "REML",
     data=Unif.len2, weights=Unif.len2$weights)
  resid.fs<-fs$residuals

  ss<-gam(score100_len~ s(smaxalu1, k=3)+ numtest_ok
  +s(sxf31sal,k=kx)+s(por_repit, k=kx)+s(por_edu_alta,k=kx)
  +s(por_prob_espac, k=kx)+s(resid.fs), method = "REML",
   data=Unif.len2, weights=Unif.len2$weights)

  cseff[i] <- ss$coeff[[2]]
  interceptSE2[i] <- ss$coeff[[1]]

# coeficientes de 1* etapa
  interceptPE2[i] <- fs$coeff[[1]]
  IV2[i] <- fs$coeff[[2]]

 }


### Second stage Confidence Intervals and standard errors

# CS effect

# 1%
q005<-quantile(cseff,0.005, na.rm=T)
q995<-quantile(cseff,0.995, na.rm=T)
q005;q995
# 5%
q025<-quantile(cseff,0.025, na.rm=T)
q975<-quantile(cseff,0.975, na.rm=T)
q025;q975

ee.cseff <- sqrt(var(cseff))
ee.cseff

# Intercept

# 1%
q005<-quantile(interceptSE2,0.005, na.rm=T)
```

```
q995<-quantile(interceptSE2,0.995, na.rm=T)
q005;q995
# 5%
q025<-quantile(interceptSE2,0.025, na.rm=T)
q975<-quantile(interceptSE2,0.975, na.rm=T)
q025;q975

ee.intSE2 <- sqrt(var(interceptSE2))
ee.intSE2


# IV (PCS)

# 1%
q005<-quantile(IV2,0.005, na.rm=T)
q995<-quantile(IV2,0.995, na.rm=T)
q005;q995
# 5%
q025<-quantile(IV2,0.025, na.rm=T)
q975<-quantile(IV2,0.975, na.rm=T)
q025;q975

ee.IV2 <- sqrt(var(IV2))
ee.IV2

# Intercept

# 1%
q005<-quantile(interceptPE2,0.005, na.rm=T)
q995<-quantile(interceptPE2,0.995, na.rm=T)
q005;q995
# 5%
q025<-quantile(interceptPE2,0.025, na.rm=T)
q975<-quantile(interceptPE2,0.975, na.rm=T)
q025;q975

ee.intPE2 <- sqrt(var(interceptPE2))
ee.intPE2
```

A.3 Appendix of Chapter 4

A.3.1 The R code to simulation reproducibility

This section presents the R (R Core Team, 2014) syntax necessary to reproduce the Monte Carlo simulation results of Chapter 4. The code is established for a particular value of the concentration parameter, specifically for $\mu^2 = 4$. Therefore, changing the value of μ^2 in the code is enough to obtain the complete set of reported results.

```
library(mvtnorm)
library(mgcv)
library(bayesIV)


### True Model Specification

num.sim=300
burnIn=5000
numSamples=30000
thin=30
numKnots=20

N<-1000
mu2=4        ### Concentration Parameter
theta <- sqrt( mu2 / N )
theta
theta1<-2
rho<-c(0.2, 0.5, 0.95)
cor=matrix(c(1,rho[2],rho[2],1),2,2)

est=function(x) { return((x-mean(x))/sqrt(var(x)))}

pb <- txtProgressBar(style=3)

REML=GCV=naive=bivDP_2=list()

### Loop for to get simulated estimation results

for (it in 1:num.sim){
if ((it/num.sim*100)%%1==0) setTxtProgressBar(pb,
 it/num.sim)
  set.seed(it)
  var=rmvnorm(N,sigma=cor)
  epsilon=var[,1]
  ve=var[,2]
  z=rnorm(N)
    xe=theta1+theta*z+ve
    x=est(xe)
    y=pnorm(x)+epsilon

  first=x~s(z)
  second=y~s(x)

  ngrid=200
  predlist= list(seq(-3,3,length.out=ngrid),
  seq(-3,3,length.out=ngrid))
  data=data.frame(y,x,z)

## Bayes IV
   bivDP_2[[it]]=try(bayesIV(first=first,second=second,
   numBurnIn=burnIn, numSamples=numSamples,scbs=F, seed=it,
   x.pred=predlist,thin=thin, data=data, mgcv=F,
   record.npcoef=F,return.samples=T, return.f=T,
   method="chol", plotpaths=F,progressBar=T,
   numKnots = numKnots) )

 if (length(bivDP_2[[it]])>1) bivDP_2[[it]]$f.sample=NULL
 if (length(bivDP_2[[it]])>1) bivDP_2[[it]]$f.pred.sample=NULL

## 2SLS REML
  prim<-try(gam(x~s(z, k=numKnots), data=data,
  method="REML"))
  r<-try(residuals(prim))
  segu<-try(gam(y~s(x,k=numKnots)+s(r,k=numKnots),
  method="REML"))
  REML[[it]]=list()
  newdata=data.frame(cbind(x,r))
  dimnames(newdata)[[2]]=c("x","r")
  REML[[it]]$data=newdata
  REML[[it]]$f=try(predict(segu,newdata=newdata,
  type="terms"))
  newdata=data.frame(cbind(seq(-3,3,length.out=ngrid),
  r[1:ngrid]))
  dimnames(newdata)[[2]]=c("x","r")
  REML[[it]]$f.pred=try(predict(segu,newdata=newdata,
  type="terms"))
  REML[[it]]$coef=coef(prim)
  REML[[it]]$Vp=prim$Vp
  REML[[it]]$coef2=coef(segu)
  REML[[it]]$Vp2=segu$Vp

## 2SLS GCV
  prim<-try(gam(x~s(z, k=numKnots), data=data))
  r<-try(residuals(prim))
  segu<-try(gam(y~s(x, k=numKnots)+s(r, k=numKnots)))
  GCV[[it]]=list()
  newdata=data.frame(cbind(x,r))
  dimnames(newdata)[[2]]=c("x","r")
  GCV[[it]]$data=newdata
  GCV[[it]]$f=try(predict(segu,newdata=newdata,
  type="terms"))
  newdata=data.frame(cbind(seq(-3,3,length.out=ngrid),
  r[1:ngrid]))
  dimnames(newdata)[[2]]=c("x","r")
  GCV[[it]]$f.pred=try(predict(segu,newdata=newdata,
  type="terms"))
  GCV[[it]]$coef=coef(prim)
  GCV[[it]]$Vp=prim$Vp
  GCV[[it]]$coef2=coef(segu)
  GCV[[it]]$Vp2=segu$Vp

## Naive: without correction
  segu2<-try(gam(y~s(x, k=numKnots)))
  naive[[it]]=list()
  newdata=data.frame(cbind(x))
  dimnames(newdata)[[2]]=c("x")
  naive[[it]]$data=newdata
  naive[[it]]$f=try(predict(segu2,newdata=newdata,
  type="terms"))
  newdata=data.frame(cbind(seq(-3,3,length.out=ngrid)))
  dimnames(newdata)[[2]]=c("x")
  naive[[it]]$f.pred=try(predict(segu2,newdata=newdata,
  type="terms"))

}

close(pb)


### Simulation results: Figure 4.1, Squared Bias,
#   Variance and MSE

center=function(x) return(x-mean(x))

# Mean estimated functions
  b22.bivDP2 = rowMeans( sapply(bivDP_2,
  function(x) return(x$f.pred[[2]])))
  b22.REML = rowMeans( sapply(REML,
  function(x) return(x$f.pred[,1])))
  b22.GCV = rowMeans( sapply(GCV,
  function(x) return(x$f.pred[,1])))
  b22.naive = rowMeans( sapply(naive,
  function(x) return(x$f.pred[,1])))


# Quantiles "estimated functions" (estimated points really)

  b22.bivDP2.q975 = apply( sapply(bivDP_2, function(x)
  return(x$f.pred[[2]])), MARGIN=1, FUN=quantile, probs=0.97)
  b22.bivDP2.q025 = apply( sapply(bivDP_2, function(x)
  return(x$f.pred[[2]])), MARGIN=1, FUN=quantile, probs=0.03)
  b22.REML.q975   = apply( sapply(REML, function(x)
  return(x$f.pred[,1])), MARGIN=1, FUN=quantile, probs=0.97)
  b22.REML.q025   = apply( sapply(REML, function(x)
  return(x$f.pred[,1])), MARGIN=1, FUN=quantile, probs=0.03)
  b22.GCV.q975   =  apply( sapply(GCV, function(x)
  return(x$f.pred[,1])), MARGIN=1, FUN=quantile, probs=0.97)
  b22.GCV.q025   =  apply( sapply(GCV, function(x)
  return(x$f.pred[,1])), MARGIN=1, FUN=quantile, probs=0.03)
  b22.naive.q975  = apply( sapply(naive, function(x)
  return(x$f.pred[,1])), MARGIN=1, FUN=quantile, probs=0.97)
  b22.naive.q025  = apply( sapply(naive, function(x)
  return(x$f.pred[,1])), MARGIN=1, FUN=quantile, probs=0.03)
```

```
 # Figure 4.1
predlist = list(seq(-3,3,length.out=ngrid))
x11()
plot( predlist[[1]],center(pnorm( predlist[[1]])),
col=1,type="l", ylim=c(-2.8,2.8), lwd=2,
ylab="Mean and 0.03-0.97 quantile range",
xlab="Endogenous variable",family="serif",
cex.axis=1.7,cex.lab=1.7)
lines(predlist[[1]],b22.bivDP2,lty=2, lwd=2, col=1)
lines(predlist[[1]],b22.REML,lty=3, lwd=2, col=1)
lines(predlist[[1]],b22.bivDP2.q975,col=1,lty=2, lwd=2)
lines(predlist[[1]],b22.bivDP2.q025,col=1,lty=2, lwd=2)
lines(predlist[[1]],b22.REML.q975,col=1,lty=3, lwd=2)
lines(predlist[[1]],b22.REML.q025,col=1, lty=3, lwd=2)
legend("bottomright",c("true  ", "BNIV   ", "2SGAM-REML   ",
"Concent. par. = 32"),lwd=c(2,2,2,2),lty=c(1,2,3,0),cex=1.4)


### Computing Squared Bias, Variance, and MSE

## Squared Bias

true.f <- center(pnorm( predlist[[1]]))
grid.jump<-(predlist[[1]][200] - predlist[[1]][1] )/ngrid
Int.SqBias.BivDP2<-sum( grid.jump*((b22.bivDP2-true.f)^2))
Int.SqBias.REML<-sum( grid.jump*( ( b22.REML-true.f )^2))
Int.SqBias.GCV<-sum( grid.jump*( ( b22.GCV - true.f )^2))
Int.SqBias.naive<-sum( grid.jump*((b22.naive-true.f )^2))
Int.SqBias.BivDP2
Int.SqBias.REML
Int.SqBias.GCV
Int.SqBias.BivDP
Int.SqBias.naive

### Integrated MSE

true.mat <- matrix(true.f, 200, num.sim)
imse_bivDP2 <- sum grid.jump*(rowMeans((sapply(bivDP_2,
function(x) return(x$f.pred[[2]])) - true.mat)^2)) )
imse_reml <- sum( grid.jump*( rowMeans((sapply(REML,
function(x) return(x$f.pred[,1])) - true.mat)^2)) )
imse_gcv <- sum( grid.jump*( rowMeans((sapply(GCV,
function(x) return(x$f.pred[,1])) - true.mat)^2)) )
imse_naive <- sum( grid.jump*( rowMeans((sapply(naive,
function(x) return(x$f.pred[,1])) - true.mat)^2)) )
imse_bivDP2
imse_reml
imse_gcv
imse_bivDP
imse_naive

### Integrated Variance

imse_bivDP2-Int.SqBias.BivDP2
imse_reml-Int.SqBias.REML
imse_gcv-Int.SqBias.GCV
imse_naive-Int.SqBias.naive
```

Appendix B

Resumen en Español

B.1 Introducción: Inferencia Causal y el Problema de Endogeneidad

Una preocupación primordial en la investigación empírica es el descubrimiento de relaciones causales. Más precisamente, si una intervención o tratamiento en particular causa, explica o motiva un efecto o resultado particular.

En presencia de experimentos aleatorios, con asignación aleatoria del tratamiento entre las unidades de estudio, es relativamente simple derivar conclusiones causales comparando el resultado promedio para los individuos en el grupo tratado con el resultado promedio en el grupo no tratado. El mecanismo de aleatorización tiende a equilibrar características observables e inobservables que hacen que los grupos sean comparables.

Por el contrario, la identificación de relaciones causales en estudios observacionales, donde el mecanismo que asigna individuos a diferentes estados de tratamiento es desconocido o no aleatorio (es decir, el análisis se realiza utilizando datos no experimentales), no es tan simple. En este caso, los individuos de ambos grupos pueden ser sistemáticamente diferentes en términos de características inobservables, lo que confunde los efectos causales del tratamiento.

Este trabajo pretende ampliar el conocimiento empírico y las posibilidades de un modelo de regresión flexible diseñado para realizar inferencia causal en aplicaciones empíricas cuando los datos provienen de un proceso de observación. Este modelo, técnicamente conocido como el Modelo de Ecuaciones Simultáneas Triangulares No Paramétricas, ayuda a mitigar el problema que surge cuando los regresores o covariables del modelo no satisfacen la condición conocida como exogeneidad, que establece que el componente aleatorio del modelo debe ser independiente en media de todos los regresores del modelo.

El análisis de regresión con datos no experimentales se usa a menudo en ciencias sociales o de la vida para inferir la existencia de una relación causal entre una variable *tratamiento* x y una variable *respuesta* y. La presencia (o ausencia) de una relación estadística entre esas variables no es una condición suficiente ni necesaria para afirmar la presencia (o ausencia) de una relación causal. Esto es así porque los valores medidos de ambas variables, de tratamiento y respuesta, se generan mediante un experimento no controlado (por ejemplo, un proceso natural o social) que en principio es desconocido por el investigador.

Por lo tanto, se requiere una especificación previa de un modelo teórico o estructural, estableciendo un vínculo causal y una dirección causal entre el tratamiento y la respuesta. Además, dicho modelo debe tener en cuenta que el tratamiento x usualmente no se asigna aleatoriamente a las unidades de población y, como resultado, puede estar relacionado con otros factores, digamos $\mathbf{z}$, que afectan sistemáticamente la respuesta y además de x.

La estimación del efecto marginal del tratamiento x sobre la respuesta y, suele ser el objetivo principal del análisis de inferencia causal basado en métodos de regresión. Se puede interpretar como: *el cambio marginal en el valor esperado de la respuesta y causado por un cambio marginal en el tratamiento x, cuando los factores adicionales en z permanecen constantes.* Para la identificación de este efecto marginal, es necesario el cumplimiento del supuesto de exogeneidad antes mecionado, el cual necesita que se eviten las siguientes situaciones:

- S.1. La existencia de un mecanismo que determina simultáneamente los valores de ambos, la respuesta y y el tratamiento x.
- S.2. La presencia de un *problema de autoselección*, que surge cuando los individuos bajo análisis pueden elegir el nivel de tratamiento x teniendo en cuenta su efecto esperado sobre el resultado y.
- S.3. Presencia del *problema de causalidad inversa*. Tal problema surge cuando no solo el tratamiento x tiene un efecto en la respuesta y pero también y tiene un impacto sobre x.

Por otro lado, desde un punto de vista empírico que involucra una aplicación a datos reales, el logro del supuesto de exogeneidad requiere la realización simultánea de los siguientes condiciones:

- C.1. Todos los factores relevantes en $\mathbf{z}$ deben medirse e incluirse en el modelo como regresores.
- C.2. La especificación del modelo de regresión debe ser suficientemente cercana al verdadero Proceso Generador de Datos.
- C.3. Todas las variables relevantes, y, x y aquellas en $\mathbf{z}$, se deben medir sin error.

Si una o más condiciones empíricas, C.1 a C.3, no están satisfechas y/o escenarios teóricos, S.1 a S.3, no se manejan correctamente, entonces se viola el supuesto de exogeneidad. Tal situación es conocida como emph problema de endogeneidad en la literatura econométrica.

Bajo estas circunstancias, los estimadores paramétricos y no paramétricos usuales para la regresión de y sobre x y $\mathbf{z}$ serán inconsistentes.

Para superar el problema de endogeneidad descrito en la sección anterior, se han desarrollado varias metodologías. Uno de los métodos pioneros en el campo es la Regresión de Variables Instrumentales (IVR), desarrollada por primera vez en el contexto del Modelo de Regresión Lineal. En general, el método de las variables instrumentales (IV) se basa en la existencia de al menos una variable adicional (es decir, el instrumento) para cada regresor endógeno en el modelo. Intuitivamente, este instrumento debe correlacionarse con su correspondiente regresor endógeno y no estar correlacionado con ningún otro factor variable en el modelo. Para lograr la identificación de la función de regresión de interés, el instrumento w debe cumplir algunas condiciones específicas:

- ID.1 El instrumento w debe tener poder explicativo sobre el tratamiento x. Esto alude al grado de asociación condicional o parcial entre el tratamiento x y el instrumento w (dados los controles adicionales $\mathbf{z}$).

- ID.2 El instrumento no debe estar relacionado con los factores no observados (que representamos con ε) que determinan la variable de resultado y. Esta condición asegura que w no es una variable explicativa relevante para y. Esto significa, junto con la condición ID.1, que el instrumento w solo afecta el resultado y a través de su efecto sobre el tratamiento endógeno x.

El objetivo de los supuestos de identificación ID.1 e ID.2 es establecer una fuente exógena de variación en el tratamiento x, a través del instrumento w. Usar la variabilidad de w de esta manera es equivalente a obtener una asignación del tratamiento x (sobre la población) que no esté influenciada por ε.

Como se señaló anteriormente, la Regresión de Variables Instrumentales (IVR) fué desarrollada por primera vez en el contexto del Modelo de Regresión Lineal. Pero cuando el modelo de interés implica una función de regresión general (no paramétrica) es necesario basarse en un estimador IV no paramétrico. Una metodología que surgió para atacar este problema de estimación es la llamada *Enfoque de Función de Control* (CFA), propuesta por primera vez por Newey, Powell, and Vella, 1999 y extendida por Pinkse, 2000, Su and Ullah, 2008 y Marra and Radice, 2011. Este enfoque CFA se basa en un modelo de ecuaciones simultáneas no paramétricas triangulares (Newey, Powell, and Vella, 1999).

En este trabajo de tesis utilizamos la metodología CFA para estimar varias funciones de regresión pero con una estructura aditiva que incluye términos no paramétricos y paramétricos. Como es conocido, dicho modelo aditivo o semi-paramétrico ayuda a evitar la maldición de la dimensionalidad que emerge en la estimación no paramétrica cuando hay un gran número de regresores.

B.2 Modelos Flexibles para la Evaluación del Tiempo Óptimo a Intervención en Pacientes con Síndrome Coronario Agudo

La intervención invasiva en pacientes con síndromes coronarios agudos sin elevación del segmento ST (SCASEST-ACS) incluye procedimientos de evaluación (como lo es el cateterismo cardíaco) y terapias como la revascularización. Los procedimientos de evaluación se implementan primero y son útiles para decidir qué terapia seguir posteriormente. La ejecución temprana de este tipo de intervenciones, generalmente antes de las 72 horas desde la asistencia del paciente, se establece como la estrategia de tratamiento recomendada, en lugar de seguir un plan conservador de administración de medicamentos.

Sin embargo, el momento óptimo para la intervención en pacientes con SCASEST sigue siendo un tema abierto al debate (Navarese et al., 2013). Algunas de las causas de estos resultados no concluyentes pueden estar relacionadas con cuestiones metodológicas, especialmente cuando se explotan los datos de registro de observación.

En estudios observacionales los procedimientos de estimación mediante análisis de regresión están expuestos a sesgos debido al problema de endogeneidad del tratamiento. Por lo tanto, los métodos de regresión basados en variables instrumentales (IV) son alternativas naturales para manejar el problema del sesgo de endogeneidad.

En la última década, varios estudios observacionales abordaron este tema, como Montalescot et al., 2005, Tricoci et al., 2007 y Sorajja et al., 2010, pero todos descuidaron el problema de endogeneidad. Una excepción notable es Ryan et al., 2005, que utiliza el día de la presentación en el hospital (fin de semana vs. día de la semana) como variable instrumental (IV) para estudiar el impacto del momento del cateterismo cardíaco y la terapia de revascularización sobre la mortalidad hospitalaria. Ellos encuentran beneficios no significativos para el cateterismo temprano, aunque no se puede excluir una importante reducción del riesgo.

Siguiendo una estrategia de identificación similar a la de Ryan et al., 2005, estudiamos el impacto del retraso en el cateterismo sobre los resultados (de mortalidad y reincidencia) para los pacientes con Insuficiencia Miocárdica con Elevación del segmento, explotando el hecho de que los pacientes ingresados los fines de semana son menos propensos a someterse a un cateterismo temprano que los pacientes ingresados durante los días laborables de la semana. Por lo tanto, empleamos esta fuente exógena de variación en el tratamiento para identificar su efecto causal en los resultados a través de modelos de regresión basados en variables instrumentales.

En contraste con el enfoque tradicional (generalmente seguido por los investigadores en esta literatura específica y empleado por Ryan et al., 2005), aquí presentamos innovaciones en dos direcciones. Por un lado, mantenemos la variable continua original *tiempo de retraso* *a cateterismo* (*TDC*) como tratamiento relevante, medida en unidades de tiempo, en lugar de especificarlo como una variable binaria (ficticia) que indica *cateterización temprana*. Por otro lado, nuestro procedimiento de inferencia causal se basa en una especificación flexible del Modelo de Ecuaciones Simultáneas Triangulares, propuesto recientemente por Marra and Radice, 2011.

La variable de resultado se define mediante una variable binaria llamada *Evento*, que indica la presencia de cualquiera de las dos situaciones: a) mortalidad por cualquier causa dede la intervención hasta los 12 meses de seguimiento y b) infracción miocárdica aguda desde la intervención hasta los 12 meses.

Nuestra base de datos incluye pacientes ingresados consecutivamente entre noviembre de 2003 y enero de 2011 al Departamento de Cardiología del Hospital Clínico de Santiago, con diagnóstico de Síndrome Coronario Agudo (SCA). Los datos demográficos y clínicos se recopilaron prospectivamente y se registraron digitalmente.

La muestra disponible incluye pacientes con SCASEST que se han sometido a cateterismo cardíaco con un retraso de entre 0 y 1000 horas. Esto nos permitió medir el retraso del tratamiento en tiempo continuo y definir la variable de tratamiento *TDC*, medida en horas.

Uno de los principales obstáculos a superar es la naturaleza endógena de *TDC*. Esto es así porque la decisión de realizar un cateterismo se basa en las características de los pacientes que pueden ser percibidas por completo por el personal médico, pero solo son observadas parcialmente por el investigador.

Si las características basales de los pacientes y los tratamientos hospitalarios habituales no difieren según cuándo se presenten los pacientes (en el día de la semana o el fin de semana) en el hospital, la condición de ser ingresado el fin de semana podría utilizarse como una IV válida para evaluar el efecto de la cateterización cardíaca.

Los pacientes ingresados el fin de semana incluyen a los que se presentaron en el hospital entre las 5 p. m. del viernes y las 3 p. m. del domingo. Todos los demás pacientes fueron considerados pacientes ingresados entre semana. Luego, las variables instrumentales se definieron como un conjunto de tres variables ficticias mutuamente excluyentes, indicando admisiones los viernes, sábados y domingos.

Finalmente, incluimos dos variables de control continuas, *Age* (que contiene la edad del paciente en años) y *Gr* (GRACE, Registro Global de Puntuaciones de riesgo de Eventos Coronarios Agudos), junto con la variable de control binaria *Fem* que indica pacientes femeninos.

Los modelos de regresión utilizados incluyen al Modelo Lineal y al Modelo Lineal Generalizado (GLM), en el caso paramétrico, y al Modelo Aditivo y al Modelo Aditivo Generalizado (GAM). Tanto en el caso del GLM como el GAM, se utiliza la función link Probit.

Pasando a los resultados empíricos, se encontró que el sesgo por endogeneidad es de un nivel importante a nivel práctico, lo que provoca que los modelos de regresión sin corrección fallen en identificar un efecto significativo del tratamiento.

Desde una perspectiva médica, los resultados apoyan la existencia de un efecto positivo no lineal de *TDC* sobre la supervivencia y el estado de salud de los pacientes. Además, el uso de un modelo flexible permite identificar un rango específico de valores para *TDC*, dede 0 hasta 30 horas aproximadamente, en el que el tratamiento muestra un efecto marginal estadísticamente significativo que varía entre 0.011 y un valor ligeramente superior a 0.

Los resultados del Modelo Aditivo, comparado con el Modelo Lineal, revelan cuán sesgado podría ser el uso descuidado de éste último, ya que implica un efecto marginal significativo igual a 0,0048 para todo el rango *TDC* (es decir, de 0 a 60 horas). En cambio, con base en el Modelo Aditivo podemos calcular un efecto marginal global de aproximadamente 0,0059 promediando en todo el rango *TDC*. Más aún, promediando solo para los valores de *TDC* para los que el efecto marginal es significativo estadísticamente (entre 0 a 34 horas), obtenemos un efecto marginal más grande de aproximadamente 0,0076. Ambas instancias demuestran una sub-estimación médicamente relevante si se utiliza el Modelo Lineal.

Los dos modelos flexibles, tanto el modelo aditivo como el GAM, aportan evidencia empírica que respalda que la cateterización temprana es una buena decisión dentro de las primeras 30 horas desde el ingreso en el hospital, y cuanto antes se realice mejores son las perspectivas de supervivencia y no reincidencia.

B.3 Identificación del Efecto del Tamaño de la Clase sobre el Rendimiento Escolar mediante Modelos de Ecuaciones Triangulates Simultáneas

Este capítulo trata sobre las ventajas de usar modelos de regresión flexibles de variables instrumentales (IV), junto con un adecuado tratamiento de observaciones atípicas, en la estimación del efecto del tamaño de la clase (definido como el número de alumnos que asisten a una clase) sobre el rendimiento escolar de los alumnos.

La estimación de este efecto suele ser problemática por el carácter endógeno del regresor de interés, es decir, el tamaño de la clase. Son varias las fuentes de dicha endogeneidad que pueden estar presentes en forma simultánea. Entre ellas se destacan la distribución no aleatoria de los alumnos en función de su buena o mala conducta en clases y la elección por parte de los padres de la escuela que ellos prefieren según su calidad educativa.

Las innovaciones respecto del procedimiento estándar consisten en, primero, usar un modelo aditivo semi-paramétrico que incorpora potenciales efectos no-lineales de las variables de control o explicativas, segundo, implementar un método de Bootstrap Ponderado para la inferencia que tenga en cuenta la estructura de clusters en las observaciones y, tercero, utilizar una estrategia adecuada para evitar el sesgo producidos por observaciones atípicas en el valor del tamaño de la clase.

El estimador utilizado se basa en el enfoque de Función de Control para la regresión con variables instrumentales. El modelo de regresión específico consiste en un sistema de ecuaciones simultáneas triangulares, desarrollado recientemente por Marra and Radice, 2011, el cual es estimado por un procedimiento en dos etapas denominado Modelo Aditivo Generalizado de Dos Etapas (2SAM). Para llevar a cabo inferencia estadística, en un contexto de datos agrupados en escuelas, se utiliza el método de remuestreo llamado Bootstrap ponedrado (Chatterjee and Bose, 2005, Chatterjee and Bose, 2000 y Bose and Chatterjee, 2002). Este método consiste en generar pesos, siendo cada uno de ellos asignado a estudiantes de una misma escuela/cluster. De este modo, cada escuela de la muestra recibe un peso aleatorio, generando una muestra Bootstrap. Luego, la estimación por 2SAM es aplicada en cada muestra Bootstrap, que es igual a la muestra original pero ponderada por los pesos aleatorios.

Finalmente, para detectar apropiadamente a las observaciones atípicas que afectan la primera etapa de estimación, proponemos aplicar una versión modificada de la estrategia descrita en Dehon, Desbordes, and Verardi, 2015. La modificación propuesta es necesaria para evitar la sub-detección de datos atípicos que ocurre al utilizar el procedimiento estándar presentado en Dehon, Desbordes, and Verardi, 2015.

El procedimiento propuesto es ilustrado mediante el análisis de una base de datos de estudiantes de sexto grado de escuelas primarias del Uruguay. Estos datos provienen de un programa de evaluación nacional que aplica una prueba estandarizada en las temáticas de literatura y matemática. En este caso se utilizan las puntuaciones de la prueba en literatura como forma de aproximar el aprovechamiento escolar de los alumnos. Para obtener una medida adecuada del tamaño de la clase, se contó el número de estudiantes que tomaron al menios una de las dos pruebas (Literatura o Matemática). Los datos también incluyen información adicional sobre características de los estudiantes, de los profesores y de la escuela, permitiendo la incorporación de variables de control a nivel de alumnos y escuela.

Las variables que se incluyen en el análisis son las siguientes:

- Puntuación en Literatura: es la calificación del alumno obtenida en la prueba de literarura.
- Tamaño de la Clase (CS): es el número de estudiantes dentro de la clase.
- Matrícula: Número total de alumnos matriculados.

- Tamaño de Clase Predicho (PCS): es la variable instrumental utilizada. La misma se basa en el cumplimiento de la regla oficial de cantidad máxima de alumnos por clase, fijada en 40 alumnos. Esta regla estipula que cuando la matrícula escolar supere los 40 alumnos, la escuela debe separar a los estudiantes en dos aulas, cuando supere los 80 alumnos debe separarlos en 3 aulas y así sucesivamente. Esta regla genera una variabilidad exógena del CS, que no debería ser modificada discrecionalmente por las escuelas, por tal motivo es considerada como una variable instrumental válida.

- Índice Socioeconómico: índice que caracteriza el contexto económico del área donde la escuela está localizada.

- Educación Alta (%): Porcentaje de alumnos en el aula para los cuales al menos uno de sus padres posee educación universitaria.

- Problemas Habitacionales (%): Porcentaje de alumnos en el aula que presentan problemas habitacionales, definido como teniendo más de dos personas por habitación en sis hogares.

- Repitentes (%): Porcentaje de alumnos en el aula que son repitentes de algún año escolar.

Enfocándonos en los resultados del análisis empírico, si se ignora el problema de endogeneidad y se aplican estimadores usuales, como Mínimos Cuadrados Ordinarios (OLS) en el modelo lineal y Máxima Verosimilitud Restringida (REML) en el modelo aditivo flexible, no se obtiene evidencia de un efecto significativo del tamaño de la clase. Para este caso, el modelo aditivo revela la existencia de efectos no-lineales para los regresores Índice Socioeconómico, Educación Alta y Problemas Habitacionales. En este mismo contexto, si se excluyen las escuelas que presentan valores atípicos de CS, los resultados obtenidos implican un aumento del efecto negativo del tamaño de clase, pero que solo resulta estadísticamente significativo utilizando el modelo aditivo flexible (aunque con un límite superior del intervalo de confianza muy cercano a cero, iagual a -0,029, que tendría poco impacto a nivel práctico).

Los resultados que se obtienen de las regresiones con la variable instrumental (TCP), tanto paramétrica con 2SLS como flexible con 2SAM, dan cuenta de un efecto negativo y significativo del tamaño de clase, cercano a -1 punto de cambio en las calificaciones por cada estudiante que se agrega a la calse. Este valor sugiere un efecto muy alto del TC a nivel práctico y la existencia de un alto grado de endogeneidad de TC.

Por otro lado, cuando las observaciones atípicas son tratadas, se obtiene un efecto estimado menor para el tamaño de clase. El modelo lineal arroja un efecto de -0,215 y el modelo aditivo muestra un efecto de -0,314. La diferencia entre estas dos estimaciones es sustancial, implicando que el modelo flexible arroja un efecto que es mayor en un 46% respecto al estimado con el modelo lineal. Adicionalmente, la significatividad estadística del efecto es débil en el caso paramétrico, donde solo al 95% de confianza el efecto resulta diferente de cero, no así para el

99% de confianza. En cambio, con el modelo flexible el efecto del TC es significativo tanto al 95% como al 99% de confianza.

Estos resultados muestran el impacto que tiene tanto la utilización de un modelo flexible como el correcto tratamiento de observaciones atípicas. En este caso puntual, la omisión de tratar las observaciones atípicas, produce una sobreestimación del efecto del TC, tanto en el modelo lineal como en el aditivo. Por otro lado, cuando se tratan adecuadamente los datos atípicos, el uso del modelo aditivo arroja una estimación significativamente mayor a nivel práctico (y con mayor evidencia estadística) respecto del resultado del modelo paramétrico.

Finalmente, el ejercicio de simulación de Monte Carlo ayuda a ilustrar sobre la necesidad e tratar los datos atípicos encontrados en algunas escuelas. Este ejercicio muestra que el uso del modelo lineal de variables instrumentales, cuando los datos atípicos están presentes, produce un elevado sesgo hacia abajo en la estimación del efecto del tamaño de clase. Por otro lado, cuando los outliers son eliminados, el sesgo desaparece al utilizar el mismo método de estimación. Lo propio ocurre cuando la estimación se realiza con un estimador robusto de variables instrumentales.

En términos de evidencia empírica, la combinación de estas tres innovaciones procedimentales ayudó a identificar un efecto del tamaño de clase que es significativo tanto a nivel estadístico como práctico. Tal efecto marginal, de -0,324 puntos, representa un valor alto en relación a los hallazgos previos de la literatura relacionada.

B.4 Estimación flexible de Modelos de Ecuaciones Simultáneas Triangulares en contextos de instrumentos débiles

La amplia familia de estimadores de regresión basados en variables instrumentales (IV) comparten una condición necesaria de identificación, a saber, la existencia de una correlación parcial "suficiente" entre los instrumentos y las variables endógenas correspondientes. El incumplimiento de esta condición se conoce como el problema de *instrumentos débiles* o *identificación débil*. En aplicaciones prácticas de estimadores estándar IV, este problema causa resultados indeseables, principalmente relacionados con sesgos de muestra finita, pérdida de precisión y falta de fiabilidad de las aproximaciones asintóticas a la distribución normal.

El problema de la identificación débil ha sido estudiado ampliamente en los últimos 20 años en el contexto de regresión paramétrica. Las principales contribuciones se pueden encontrar en cite bound1995problems, Staiger and Stock, 1997, Stock and Wright, 2000, Kleibergen, 2002, Stock and Yogo, 2005, Newey and Windmeijer, 2009 y Andrews and Cheng, 2012, todas ellos pertenecientes a la literatura frecuentista. Recientemente, se han propuesto enfoques bayesianos para la estimación IV, que funcionan mejor que las alternativas frecuentistas tradicionales (en términos de sesgo y cobertura del intervalo de confianza) en ciertos escenarios caracterizados por una identificación débil (ver Burgess and Thompson, 2012 y Conley et al., 2008).

En el contexto de regresión no paramétrica, se han realizado varios esfuerzos para diseñar un método IV confiable. A pesar de estas contribuciones, el estudio del problema de los instrumentos débiles en este contexto ha sido en descuidado. Una excepción, dentro del enfoque frecuentista, es el trabajo en progreso provisto por Han, 2014, que define el problema de identificación débil para el Modelo de Ecuaciones Simultáneas Triangulares flexible, y propone un método de estimación de serie penalizado que alivia el efecto de los instrumentos débiles.

En este capítulo se propone un método IV bayesiano no paramétrico para modelos de ecuaciones triangulares con una variable endógena (BNIV), publicado recientemente en Wiesenfarth et al., 2014, el cual parece resultar competitivo en términos de aliviar el efecto de identificación débil. Este método bayesiano tiene una ventaja sobre el método frecuentista presentado en Han, 2014 ya que realiza la estimación de todos los parámetros de ajuste necesarios (incluidos los parámetros de suavizado) a partir de los datos disponibles.

Mediante dos ejercicios de simulación se establece una comparación, en el contexto de un escenario de identificación débil, entre el método IV bayesiano propuesto y una alternativa frecuentista conocida como Modelo Aditivo Generalizado de Dos Etapas (2SGAM) introducido por Marra and Radice, 2011. Ambas alternativas bayesiana y frecuentista son comparables en varios aspectos, como ser la selección automática de los parámetros de suavizado y el uso de splines para la especificación de las funciones de base.

La comparación entre los enfoques 2SGAM y BNIV, en el contexto de escenarios de identificación débiles, se lleva a cabo ejecutando dos simulaciones de Monte Carlo basados en proceso de generación de datos utilizado en Han, 2014. Para las simulaciones se establece la secuencia de valores $\{4, 10, 12, 16, 32, 64, 256\}$ para el parámetro de concentración, que incluye el umbral de 10 (que define si un instrumento es débil o fuerte en el modelo lineal o paramétrico) y valores que van desde instrumentos débiles a fuertes.

Para grados decrecientes en la fuerza del instrumento, hasta un límite inferior de 10 para el parámetro de concentración, el sesgo al cuadrado obtenido mantiene valores relativamente bajos y estables para el estimador 2SGAM, pero la varianza aumenta en forma acelerada.

En términos de Error Cuadrático Medio Integrado (IMSE), y para un parámetro de concentración menor o igual a 10, 2SGAM presenta un rendimiento peor que el estimador naive (sin corrección de endogeneidad). Esto revela la necesidad de establecer un valor más alto para el parámetro de concentración como el nuevo umbral para caracterizar escenarios de identificación débiles no paramétricos, en lugar del umbral de 10 establecido en el caso paramétrico. Para valores de 12, 16 y 32 del parámetro de concentración, los IMSE del estimador 2SGAM representan aproximadamente 60 %, 40 % y 15 % del IMSE de estimador naive, respectivamente. Estos resultados sugieren que el nuevo umbral para el parámetro de concentración se podría especificar como un valor entre 12 y 16.

Comparado con el estimador 2SGAM y para un parámetro de concentración que va de 10 a 16, el método BNIV muestra un sesgo cuadrado igual o más pequeño y una varianza menor con un comportamiento notablemente más suave. Esto permite que el estimador BNIV presente un MSE relativo (respecto del estimador naive) del 45 %, 40 % y 30 % para valores del parámetro de concentración igual a 10, 12 y 16, respectivamente. Este resultado sugiere que el umbral de identificación débil (para el caso paramétrico) puede ser válido en el caso de BNIV. Para un parámetro de concentración igual a 10, el BNIV supera al 2SGAM en términos de sesgo y varianza.

Las consideraciones anteriores apoyan la idea de que el estimador BNIV debe ser la opción preferida cuando instrumentos débiles están presentes (en particular, cuando el parámetro de concentración esté entre 10 y 16), principalmente por su capacidad de mitigar la variabilidad de las estimaciones inducida por el problema de identificación débil.

Finalment, es destacable que la mayor eficiencia antes mencionada del estimador BNIV no se obtiene a costa de un sesgo mayor, de hecho, el sesgo de BNIV mantiene valores estables que son, en general, más pequeños que el sesgo del estimador 2SGAM.

Bibliography

Ai, C. and X. Chen (2003). “Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions”. In: *Econometrica* 71.6, pp. 1795–1843.

Amemiya, T. (1978). “The Estimation of a Simultaneous Equation Generalized Probit Model”. In: *Econometrica* 46.5, pp. 1193–1205.

Andrews, D. W. and X. Cheng (2012). “Estimation and Inference With Weak, Semi-Strong, and Strong Identification”. In: *Econometrica* 80.5, pp. 2153–2211.

Angrist, J. D. and G. W. Imbens (1995). “Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity”. In: *Journal of the American Statistical Association* 90.430, pp. 431–442.

Angrist, J. D. and V. Lavy (1999). “Using Maimonides’ Rule to Estimate the Effect of Class Size on Scholastic Achievement*”. In: *The Quarterly Journal of Economics* 114.2, pp. 533–575.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton University Press.

Antoniak, C. E. (1974). “Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems”. In: *The annals of statistics*, pp. 1152–1174.

Baiocchi, M., J. Cheng, and D. S. Small (2014). “Instrumental Variable Methods for Causal Inference”. In: *Statistics in Medicine* 33.13, pp. 2297–2340.

Barbe, P. and P. Bertail (1995). *The Weighted Bootstrap*. Vol. 98. Springer Science & Business Media.

Baum, C. F., M. E. Schaffer, and S. Stillman (2007). “Enhanced Routines for Instrumental Variables/GMM Estimation and Testing”. In: *Stata Journal* 7.4, pp. 465–506.

Baum, C. F., M. E. Schaffer, S. Stillman, et al. (2003). “Instrumental Variables and GMM: Estimation and Testing”. In: *Stata journal* 3.1, pp. 1–31.

Bonesrønning, H. (2003). “Class size effects on student achievement in Norway: Patterns and explanations”. In: *Southern Economic Journal*, pp. 952–965.

Bose, A. and S. Chatterjee (2002). "Comparison of Bootstrap and Jackknife Variance Estimators in Linear Regression: Second Order Results". In: *Statistica Sinica* 12, pp. 575–598.

Bound, J., D. A. Jaeger, and R. M. Baker (1995). "Problems with Instrumental Variables Estimation when the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak". In: *Journal of the American Statistical Association* 90.430, pp. 443–450.

Bowden, R. J. and D. A. Turkington (1990). *Instrumental Variables*. Vol. 8. Cambridge University Press.

Burgess, S. and S. G. Thompson (2012). "Improving Bias and Coverage in Instrumental Variable Analysis with Weak Instruments for Continuous and Binary Outcomes". In: *Statistics in Medicine* 31.15, pp. 1582–1600.

Cadarso-Suárez, C. et al. (2005). "Non-parametric Estimation of the Odds Ratios for Continuous Exposures using Generalized Additive Models with an Unknown Link Function". In: *Statistics in Medicine* 24.8, pp. 1169–1184.

Chatterjee, S. and A. Bose (2000). "Variance Estimation in High Dimensional Regression Models". In: *Statistica Sinica* 10, pp. 497–515.

Chatterjee, S., A. Bose, et al. (2005). "Generalized Bootstrap for Estimating Equations". In: *The Annals of Statistics* 33.1, pp. 414–436.

Chen, X. and D. Pouzo (2009). "Efficient estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals". In: *Journal of Econometrics* 152.1, pp. 46–60.

Cheng, G., Z. Yu, and J. Z. Huang (2013). "The Cluster Bootstrap Consistency in Generalized Estimating Equations". In: *Journal of Multivariate Analysis* 115, pp. 33–47.

Chernozhukov, V., I. Fernández-Val, and A. E. Kowalski (2014). "Quantile Regression with Censoring and Endogeneity". In: *Journal of Econometrics* 186.1, pp. 201–221.

Conley, T. G. et al. (2008). "A Semi-Parametric Bayesian Approach to the Instrumental Variable Problem". In: *Journal of Econometrics* 144.1, pp. 276–305.

Cordero, J. M., V. Cristóbal, and D. Santín (2017). "Causal Inference on Education Policies: A Survey of Empirical Studies Using PISA, TIMSS and PIRLS". In: *Journal of Economic Surveys*.

Darolles, S. et al. (2011). "Nonparametric Instrumental Regression". In: *Econometrica* 79.5, pp. 1541–1565.

Dehon, C., R. Desbordes, and V. Verardi (2015). "The pitfalls of ignoring outliers in instrumental variables estimations: an application to the deep determinants of development". In:

Empirical Economic and Financial Research: Theory, Methods and Practice. Springer International Publishing, pp. 195–213.

Donoho, D. L. (1982). *Breakdown properties of multivariate location estimators*. Tech. rep. Technical report, Harvard University, Boston. URL http://www-stat. stanford. edu/˜ donoho/Reports/Old pdf.

Dreze, J. H. and J.-F. Richard (1983). “Bayesian Analysis of Simultaneous Equation Systems”. In: *Handbook of Econometrics* 1, pp. 517–598.

Eilers, P. H. and B. D. Marx (1996). “Flexible Smoothing with B-splines and Penalties”. In: *Statistical Science*, pp. 89–102.

Finlay, K. and L. M. Magnusson (2014). *Bootstrap Methods for Inference with Cluster-sample IV Models*. University of Western Australia, Business School, Economics.

Freue, G. V. C., H. Ortiz-Molina, and R. H. Zamar (2013). “A natural robustification of the ordinary instrumental variables estimator”. In: *Biometrics* 69.3, pp. 641–650.

Gary-Bobo, R. J. and M.-B. Mahjoub (2013). “Estimation of Class-Size Effects, Using" Maimonides’ Rule" and Other Instruments: the Case of French Junior High Schools”. In: *Annals of Economics and Statistics*, pp. 193–225.

Gronau, R. (1973). *Wage Comparisons -A Selectivity Bias*. Working Paper 13. National Bureau of Economic Research.

Haavelmo, T. (1943). “The Statistical Implications of a System of Simultaneous Equations”. In: *Econometrica* 11.1, pp. 1–12.

Hahn, J., P. Todd, and W. Van der Klaauw (2001). “Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design”. In: *Econometrica* 69.1, pp. 201–209.

Hall, P., J. L. Horowitz, et al. (2005). “Nonparametric Methods for Inference in the Presence of Instrumental Variables”. In: *The Annals of Statistics* 33.6, pp. 2904–2929.

Han, S. (2014). *Nonparametric Estimation of Triangular Simultaneous Equations Models under Weak Identification*. Tech. rep. The University of Texas at Austin, Department of Economics.

Hanushek, E. A. and L. Woessmann (2017). “School resources and student achievement: a review of cross-country economic research”. In: *Cognitive Abilities and Educational Outcomes: A Festschrift in Honour of Jan-Eric Gustafsson*. Springer International Publishing, pp. 149–171.

Hastie, T. and R. Tibshirani (1986). “Generalized Additive Models”. In: *Statistical Science* 1.3, pp. 297–310.

Hausman, J. A. (1978). “Specification Tests in Econometrics”. In: *Econometrica* 46.6, pp. 1251–1271.

Hausman, J. A. et al. (1983). “Specification and estimation of simultaneous equation models”. In: *Handbook of econometrics* 1.1, pp. 391–448.

Heckman, J. J. (1974). “Shadow Prices, Market Wages, and Labor Supply”. In: *Econometrica* 42.4, pp. 679–694.

— (1978). “Dummy Endogenous Variables in a Simultaneous Equation System”. In: *Econometrica* 46.4, pp. 931–959.

— (1990). “Varieties of Selection Bias”. In: *The American Economic Review* 80.2, pp. 313–318.

— (2005). “The scientific model of causality”. In: *Sociological methodology* 35.1, pp. 1–97.

— (2008). “Econometric Causality”. In: *International Statistical Review* 76.1, pp. 1–27.

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* 6, pp. 4779–4874.

Horowitz, J. L. (2014). “Adaptive Nonparametric Instrumental Variables Estimation: Empirical Choice of the Regularization Parameter”. In: *Journal of Econometrics* 180.2, pp. 158–173.

Horowitz, J. L. and E. Mammen (June 2011). “Oracle-efficient Nonparametric Estimation of an Additive Model with an Unknown Link Function”. In: *Econometric Theory* 27 (Special Issue 03), pp. 582–608.

Jara, A. et al. (2011). “DPpackage: Bayesian Semi-and Nonparametric Modeling in R”. In: *Journal of Statistical Software* 40.5, p. 1.

Kleibergen, F. (2002). “Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression”. In: *Econometrica* 70.5, pp. 1781–1803.

Kleibergen, F. and H. K. Van Dijk (1998). “Bayesian Simultaneous Equations Analysis using Reduced Rank Structures”. In: *Econometric Theory* 14.06, pp. 701–743.

Konstantopoulos, S. and T. Shen (2016). “Class size effects on mathematics achievement in Cyprus: evidence from TIMSS”. In: *Educational Research and Evaluation* 22.1-2, pp. 86–109.

Lang, S. and A. Brezger (2004). “Bayesian P-splines”. In: *Journal of Computational and Graphical Statistics* 13.1, pp. 183–212.

Lazear, E. P. (2001). “Educational Production”. In: *The Quarterly Journal of Economics* 116.3, pp. 777–803.

Leuven, E., H. Oosterbeek, and M. Rønning (2008). “Quasi-experimental estimates of the effect of class size on achievement in Norway”. In: *The Scandinavian Journal of Economics* 110.4, pp. 663–693.

Li, W. and S. Konstantopoulos (2016). “Class Size Effects on Fourth-Grade Mathematics Achievement: Evidence From TIMSS 2011”. In: *Journal of Research on Educational Effectiveness* 9.4, pp. 503–530.

Ma, S. and M. R. Kosorok (2005). “Robust Semiparametric M-estimation and the Weighted Bootstrap”. In: *Journal of Multivariate Analysis* 96.1, pp. 190–217.

Maddala, G. S. (1986). *Limited-dependent and Qualitative Variables in Econometrics*. 3. Cambridge University Press.

Marra, G. and R. Radice (2011). “A Flexible Instrumental Variable Approach”. In: *Statistical Modelling* 11.6, pp. 581–603.

Marra, G. and S. N. Wood (2012). “Coverage Properties of Confidence Intervals for Generalized Additive Model Components”. In: *Scandinavian Journal of Statistics* 39.1, pp. 53–74.

McCullagh, P. and J. A. Nelder (1989). *Generalized linear models*. Vol. 37. CRC press.

Mishel, L. R. et al. (2002). *The Class Size Debate*. Economic Policy Institute.

Montalescot, G. et al. (2005). “Relation of Timing of Cardiac Catheterization to Outcomes in Patients with Non–ST-segment Elevation Myocardial Infarction or Unstable Angina Pectoris Enrolled in the Multinational Global Registry of Acute Coronary Events”. In: *The American Journal of Cardiology* 95.12, pp. 1397–1403.

Moulton, B. R. (1986). “Random Group Effects and the Precision of Regression Estimates”. In: *Journal of Econometrics* 32.3, pp. 385–397.

Navarese, E. P. et al. (2013). “Optimal Timing of Coronary Invasive Strategy in Non–ST-Segment Elevation Acute Coronary SyndromesA Systematic Review and Meta-analysis”. In: *Annals of Internal Medicine* 158.4, pp. 261–270.

Nelder, J. A. and R. W. M. Wedderburn (1972). “Generalized Linear Models”. In: *Journal of the Royal Statistical Society. Series A (General)* 135.3, pp. 370–384.

Newey, W. K. (1987). “Efficient Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables”. In: *Journal of Econometrics* 36.3, pp. 231–250.

Newey, W. K. and J. L. Powell (2003). “Instrumental Variable Estimation of Nonparametric Models”. In: *Econometrica* 71.5, pp. 1565–1578.

Newey, W. K., J. L. Powell, and F. Vella (1999). “Nonparametric Estimation of Triangular Simultaneous Equations Models”. In: *Econometrica* 67.3, pp. 565–603.

Newey, W. K. and F. Windmeijer (2009). “Generalized Method of Moments with Many Weak Moment Conditions”. In: *Econometrica* 77.3, pp. 687–719.

Newton, M. A. and A. E. Raftery (1994). “Approximate Bayesian Inference with the Weighted Likelihood Bootstrap”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 56.1, pp. 3–48.

Pinkse, J. (2000). “Nonparametric Two-step Regression Estimation when Regressors and Error are Dependent”. In: *Canadian Journal of Statistics* 28.2, pp. 289–300.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Reiss, P. T. and R Todd Ogden (2009). “Smoothing Parameter Selection for a Class of Semiparametric Linear Models”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.2, pp. 505–523.

Roca-Pardiñas, J. et al. (2004). “Predicting Binary Time Series of SO2 using Generalized Additive Models with Unknown Link Function”. In: *Environmetrics* 15.7, pp. 729–742.

Rosenbaum, P. R. (1984). “From Association to Causation in Observational Studies: The Role of Tests of Strongly Ignorable Treatment Assignment”. In: *Journal of the American Statistical Association* 79.385, pp. 41–48.

— (2002). *Observational Studies*. Springer.

Rubin, D. B. et al. (1981). “The Bayesian Bootstrap”. In: *The Annals of Statistics* 9.1, pp. 130–134.

Ryan, J. W. et al. (2005). “Optimal Timing of Intervention in Non–ST-Segment Elevation Acute Coronary Syndromes: Insights From the CRUSADE (Can Rapid risk stratification of Unstable angina patients Suppress ADverse outcomes with Early implementation of the ACC/AHA guidelines) Registry”. In: *Circulation* 112.20, pp. 3049–3057.

Shaw, P., M. A. Cohen, and T. Chen (2016). “Nonparametric Instrumental Variable Estimation in Practice”. In: *Journal of Econometric Methods* 5.1, pp. 153–177.

Shore-Sheppard, L. (1996). *The Precision of Instrumental Variables Estimates with Grouped Data*. Vol. 374. Industrial Relations Section, Princeton University.

Sorajja, P. et al. (2010). “Impact of Delay to Angioplasty in Patients With Acute Coronary Syndromes Undergoing Invasive Management: Analysis From the ACUITY (Acute Catheterization and Urgent Intervention Triage strategy) Trial”. In: *Journal of the American College of Cardiology* 55.14, pp. 1416–1424.

Stahel, W. A. (1981). “Robuste schätzungen: infinitesimale optimalität und schätzungen von kovarianzmatrizen”. PhD thesis. ETH Zurich.

Staiger, D. and J. H. Stock (1997). “Instrumental Variables Regression with Weak Instruments”. In: *Econometrica* 65.3, pp. 557–586.

Stock, J. H. and J. H. Wright (2000). “GMM with Weak Identification”. In: *Econometrica* 68.5, pp. 1055–1096.

Stock, J. H., J. H. Wright, and M. Yogo (2002). “A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments”. In: *Journal of Business & Economic Statistics* 20.4, pp. 518–529.

Stock, J. H. and M. Yogo (2005). “Testing for Weak Instruments in Linear IV Regression”. In: *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*.

Su, L. and A. Ullah (2008). “Local Polynomial Estimation of Nonparametric Simultaneous Equations Models”. In: *Journal of Econometrics* 144.1, pp. 193–218.

Tricoci, P. et al. (2007). “Time to Coronary Angiography and Outcomes Among Patients With High-Risk Non–ST-Segment–Elevation Acute Coronary Syndromes: Results From the SYNERGY Trial”. In: *Circulation* 116.23, pp. 2669–2677.

Tutz, G. and S. Petry (2013). *Generalized Additive Models with Unknown Link Function Including Variable Selection*. Tech. rep. 145.

Urquiola, M. (2006). “Identifying Class Size Effects in Developing Countries: Evidence from Rural Bolivia”. In: *Review of Economics and statistics* 88.1, pp. 171–177.

White, H. (2014). *Asymptotic Theory for Econometricians*. Academic Press.

Wiesenfarth, M. et al. (2014). “Bayesian Nonparametric Instrumental Variables Regression Based on Penalized Splines and Dirichlet Process Mixtures”. In: *Journal of Business & Economic Statistics* 32.3, pp. 468–482.

Winship, C. and S. L. Morgan (1999). “The Estimation of Causal Effects from Observational Data”. In: *Annual Review of Sociology* 25, pp. 659–707.

Woessmann, L. (2005). “Educational production in Europe”. In: *Economic policy* 20.43, pp. 446–504.

Woessmann, L. (2006). "International Evidence on Expenditures and Class Size: A Review". In: *Brookings Papers on Education Policy* 9, pp. 245–272.

Wood, S. (2006a). *Generalized Additive Models: an Introduction with R*. CRC Press.

Wood, S. N. (2003). "Thin Plate Regression Splines". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.1, pp. 95–114.

— (2004). "Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models". In: *Journal of the American Statistical Association* 99.467, pp. 673–686.

— (2006b). "On Confidence Intervals for Generalized Additive Models based on Penalized Regression Splines". In: *Australian & New Zealand Journal of Statistics* 48.4, pp. 445–464.

— (2011). "Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.1, pp. 3–36.

— (2013). "On P-values for Smooth Components of an Extended Generalized Additive Model". In: *Biometrika* 100.1, pp. 221–228.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT press.

Wright, P. G. et al. (1928). *Tariff on Animal and Vegetable Oils*. The Macmillan Co.

Zanin, L., R. Radice, and G. Marra (2014). "A Comparison of Approaches for Estimating the Effect of Women's Education on the Probability of using Modern Contraceptive Methods in Malawi". In: *The Social Science Journal* 51.3, pp. 361–367.

Zhelonkin, M., M. G. Genton, and E. Ronchetti (2012). "On the robustness of two-stage estimators". In: *Statistics & Probability Letters* 82.4, pp. 726–732.