Abstract

Assessing the heterogeneous causal effects of endogenous variables is of a strong interest in econometrics. In this context, the varying-coefficient models have a natural potential to model the heterogeneous effects as non-parametric functions. On the one hand, we can introduce the use of instruments in the varying-coefficient model framework with a control function approach. On the other hand, the usual conditional independence condition of instruments assumed in the fully non-parametric case can be relaxed to express the coefficient function with respect to the instruments. We illustrate our approach by estimating the effect of the American public health insurance Medicaid on the household savings by using an adaptation of an available and effective R package implementation.

Reference


URN : urn:nbn:ch:unige-1044426
DOI : 10.13097/archive-ouverte/unige:104442

Available at:
http://archive-ouverte.unige.ch/unige:104442

Disclaimer: layout of this document may differ from the published version.
Heterogeneous Coefficient Identification and Estimation in Econometric Models

by

Raoul THELER

A dissertation submitted to the
Geneva School of Economics and Management,
University of Geneva, Switzerland,
in fulfilment of the requirements for the degree of
Ph.D. in Econometrics.

Members of the dissertation committee:

Prof. Stefan Andreas SPERLICH, Advisor, University of Geneva
 Prof. Jaya KRISHNAKUMAR, Chair, University of Geneva
 Prof. Juan Manuel RODRIGUEZ-POO, University of Cantabria

Thesis No. 50

April, 2018
Acknowledgement

A Ph.D. dissertation is a collective work of which the external contributions cannot be acknowledged in such a short space. Among the people who helped me the most, I would like to express my sincere gratefulness to my advisor, Professor Stefan Sperlich, who gave me the chance to experience many aspects of the researcher task.

I would also like to thank my committee members: Professor Jaya Krishnakumar, Professor Juan Rodriguez-Poo, for their attention and valuable comments as well as all the other professors who helped and supported me during my research including Professor Gérard Antille, Professor Martin Huber, and Professor Christian Ghiglino.

I also would like to express my gratitude to my Ph.D. colleagues and dear friends: Annick Valleau, Yann Miquel, Cenni Najy, Younes Ahouga, Mariem Nouar, and Giacomo Benini. Finally, I especially thank my family for all the sacrifices they have made on my behalf for all these years.
Abstract

The two last decades have been characterized by a phenomenal progress in IT technologies. The relating computational power and the abundance of information of the recent "big-data" environment enable a higher informative complexity in statistical modeling. As a result, many extensions and adaptations of the initial parametric statistical theory have been required. The relating data-driven flexibility in modeling relax many potentially over-assumed parametric structures of the theoretical frameworks.

In econometrics, this shift to a more empirical oriented approach is well illustrated by the treatment effect and policy evaluation literature that is based on a non-parametric framework. The latter is more flexible than the standard parametric approach of the structural models that are potentially over-assumed. This flexibility enables to model the individual heterogeneity that has been drastically underestimated in the initial yet dominant linear models. However, despite the continuous increase of the sample sizes, a full non-parametric is subject to the so-called curse of dimensionality well know in statistics. Therefore, a trade-off between non-parametric flexibility and parametric assumptions has to be found. The semi-parametric modeling offers a useful alternative yet seldom implemented in econometrics. The present dissertation aims to investigate the validity and pertinence of semi-parametric assumptions in theoretical econometrics when the individual heterogeneity is modeled with non-parametric functions.
The heterogeneity of causal effects is of a strong interest in econometrics. In this context, the slope coefficients of the widely used linear models are considered to be the average returns of individual decisions. First, the inadequacy of this standard approach is illustrated. Then, it is shown why semi-parametric varying-coefficient models have a strong natural potential to model heterogeneity as in many interesting regression problems. Moreover, it is straightforward to develop alternative IV specifications in the varying-coefficient model framework. The corresponding modeling and implementation software facilities that are available in R are studied in detail. On the other hand, a fully non-parametric specification liked considered in treatment effect literature requires strong conditional independence assumptions that are often over-assumed. In this context, the semi-parametric representation that enables to assess the effects when these requirements are not fulfilled is valuable. We illustrate our approach by estimating the effect of the American public health insurance Medicaid on the household savings.
Contents

Acknowledgement i

Abstract iii

Introduction 1

1 Varying Coefficient Models Revisited: an Econometric View 5
  1.1 The Causality Problem in the Presence of Heterogeneous Returns 5
  1.2 Triangular Varying Coefficient Models with Instruments 15
  1.3 An Example 21

2 Modeling Heterogeneity: A Praise for Varying-coefficient Models in Causal Analysis 29
  2.1 Introduction 29
  2.2 Problems with linear OLS and IV estimation 34
  2.3 The varying-coefficient model for heterogeneous effects 40
    2.3.1 When \( u \) is mean-independent from \( X \) given \( W \) 41
2.3.2 Dependence between \( u \) and \( X \) given \( W \) but existence of an instrument \( Z \) ........................................ 42

2.4 Estimating varying-coefficient models in practice .......................... 44

2.5 Conclusion ........................................................................... 60

3 On Semi-parametric Models for Evaluating Endogenous Effects Correlated with Instruments through Unobservable Abilities 67

3.1 Introduction ................................................................. 67

3.2 Unobservable Abilities and Their Influence on The Instruments . 70

3.3 Semi-parametric Representations and Identification ................. 72

3.4 Estimation and Confidence Intervals ................................. 78

3.5 The Empirical Illustration .............................................. 82

3.6 Conclusion ........................................................................ 87

Appendices ............................................................................. 103

A Computer Script of Chapter 2 .............................................. 105
A ma très chère famille: Ma mère et mon frère.
Introduction

In econometrics, the easiest way to represent heterogeneity is probably to consider an individual heterogeneous coefficient sampling model as follows:

\[ Y_i = \alpha_i + X_i^T \beta_i, \quad i = 1, \ldots, n. \]

This model is similar to the widely spread multivariate linear model up to the coefficients that here vary with individuals. Econometricians usually aim to estimate the overall average coefficients, namely \( E[\beta]. \) As it is well established in binary treatment effect and policy evaluation literature, when \( X \) is correlated with \( \beta, \) the usual Ordinary Least Square and the Instrumental Variable methods fail to estimate consistently the average effects. This issue occurs, when individuals select themselves, and, goes much beyond the scope of the binary treatment case that is usually considered in impact evaluation although a possible remedy shares a similar principle; average effect parameters on sub-populations have to be considered. However, the fully non-parametric approach of the treatment effect and policy evaluation econometrics has defects. First, the strong conditional independence assumptions, required for the identification of average effect parameters, are often not fulfilled in practice. Second, the curse of dimensionality impedes a reliable point-wise estimation of the non-parametric functions.

We focus on semi-parametric models that enable us to deal with spurious associations between \( X \) and \( \beta \) due to self-selection. The related impact has been underestimated in the general cross-sectional linear models though yielding to potentially
very high finite sample biases as emphasized by simulation studies in Chapter 2. As a result, we model the effects on sub-populations as a function of variables that control for these spurious association. On this purpose, we semi-parametrize Model () as follows:

\[ Y_i = \alpha + X_i^T \beta(W_i, V_i) + \varepsilon_i, \quad E[\varepsilon_i] = 0 \quad i = 1,..,n. \quad (1) \]

This is a particular case of the varying-coefficient models where \( \beta \) is a non-parametric function and \((W, V)\) are effect modifiers. In this case, the average effects depend on observable characteristics; \( W \) and unobservable ones \( V \). As \( V \) is unobserved, the easiest way to consider \( \beta(W, V) \) is to define it in terms of projection: \( \beta(W, V) = g(W) + U \) with \( g(W) := E[\beta|W] \) such that \( E[U|W] = 0 \) by construction.

Then, Equation 1 becomes:

\[ Y_i = \alpha + X_i^T g(W_i) + X_i^T U_i + \varepsilon_i, \quad E[\varepsilon_i] = 0, \quad i = 1,..,n. \quad (2) \]

Thus, \( g \) provides the average effect for a sub-population spanned by observable effect modifier while \( U \) appears as a nuisance rule out in the composite error term: \( \varepsilon := X^T U + \varepsilon \).

If \( U \) and \( \varepsilon \) are mean-independent of \( X \) given \( W \), then the composite error term \( \varepsilon \) is mean-independent of \((X, W)\) and we end up with a trivial varying-coefficient model. However, in many relevant cases, some regressors violate these identification conditions to estimate \( g(\cdot) \) consistently.

Chapter 1 and 2 investigate ways to deal with endogeneity when \( E[\varepsilon|X, W] \neq 0 \) by considering respectively \( \varepsilon \) and \( U \). Chapter 1 also considers economic situations that are naturally suitable to be represented by Model 2 while Chapter 2 focuses on the statistical implementation with Monte Carlo simulations on different specifications of Model (2) to compare the most known R packages. Chapter 3 generalizes to non-additive \( \varepsilon \) and \( U \).
Chapter 1 considers the case of the usual endogeneity issue related to \( \varepsilon_i \) only such that \( E[\varepsilon_i|X_i, W_i] \neq 0 \) as follows:

\[
E[Y_i|W_i, V_i] = \alpha + X_i^T g(W_i) + \lambda(V_i) , \\
X_i = h(W_i, Z_i) + V_i , \quad i = 1, \ldots, n ,
\]

such that \( \lambda(V_i) := E[\varepsilon_i|V_i, W_i, X_i] \) and \( h(W_i, Z_i) := E[X_i|W_i, Z_i] \), where \( \lambda \) and \( h \) are non-parametric functions that model conditional expectations as follows: \( \lambda(V_i) := E[U_i|V_i] \) and \( h(Z_i) := E[X_i|W_i, Z_i] \). This specification is the direct adaptation of the model of the non-parametric model with additive control function \( \lambda \) to a varying-coefficient model.

Chapter 2 considers the case of an endogeneity related to \( U_N \) that enables to express Model (2) as follows

\[
E[Y_i|W_i, V_i, X_i] = X_i^T g(W_i) + X_i^T \lambda(V_i) , \\
X_i = h(Z_i) + V_i , \quad i = 1, \ldots, n ,
\]

The question of the inclusion of the effect modifier \( W \) in \( h \) as usual covariates is addressed in Chapter 3.

Chapter 3 considers a more general Specification of Model (1). Indeed, Specification (3) and (4) are based on the additive separability of \( W \) and \( V \) which is not the case in general. On this purpose, we consider the endogenous variable \( S \) as a continuous treatment (variable of interest) and obtain the following specification:

\[
E[Y_i|S_i, V_i, W_i, X_i] = \alpha(V_i, W_i, X_i) + \beta(V_i, W_i, X_i) S_i , \\
S_i = g(X_i, Z_i) - V_i , \quad i = 1, \ldots, n ,
\]

such that \( g(W_i, Z_i) := E[S_i|W_i, Z_i] \). \( W \) are covariates that are part of the information set of each individual when she takes her decision to participate and \( X \) covariates that are observed when the outcome is observed.
This specification is coherent with the Generalized Roy Model extended with unobservable abilities that violate the conditional independence assumption of potential outcomes such that $\alpha, \beta \not\perp Z | (W, X)$. In this case, the usual identification does not apply. As a result, we rather identify the following parameter:

$$E[\beta_i | g(X_i, Z_i) - S_i, X_i, W_i],$$

that has a policy relevance when $Z_i$ is under the control of the policymaker. We illustrate Specification (5) with the Medicaid data. We find a deterrent effect of the health insurance on the savings that is higher than the figures found in the related literature.
Chapter 1

Varying Coefficient Models
Revisited: an Econometric View

1.1 The Causality Problem in the Presence of Heterogeneous Returns

Disentangling causality from correlation is one of the fundamental problems of data analysis. Every time the experimental methodology – typical in some hard sciences – is not applicable, it becomes almost impossible to separate causality from observed correlations using non-simulated data. The only available alternative is to find a set of non testable assumptions that allow to express the causal links as parameters or as functions, and to subsequently find consistent estimators for the conditional moments or distributions that describe the parameters (or functions) of interest. In particular, consider a response $Y$ to be regressed on an explanatory variable $W$. The assumption that transforms a simple (cor)relation into a causal

effect of $W$ on $Y$, is often called 'exogeneity'.

**Definition 1.** A variable $W$ is weakly exogenous for the parameter of interest $\psi$, if and only if there exists a re-parametrization $\lambda$ for the joint density with parameter $\lambda = (\lambda_1, \lambda_2)$ such that

1. $f(y, w|\lambda_1, \lambda_2) = f(y|w; \lambda_1)f(w|\lambda_2)$.
2. $\psi$ depends on $\lambda_1$ only.
3. $(\lambda_1, \lambda_2)$ are variation free, i.e.: $(\lambda_1, \lambda_2) \in (\Lambda_1 \times \Lambda_2)$ for two given sets $\Lambda_1$, $\Lambda_2$.

The factorization presented in Definition 1 implies that the conditional density of $Y$ given $W$ is fully characterized by $\lambda_1$, while $\lambda_2$ is a so-called nuisance parameter (Engle, Hendry and Richard, 1983). In other words, if the causal impact of $W$ on $Y$ is the objective of interest, then the characterization of the distribution of $W$ is unimportant. This convenient factorization allows to focus exclusively on the relationship between $Y$ and $W$ ignoring all the other associations.

In econometrics, an outcome equation that describes the relationship between $Y$ and $W$ often has a less restrictive moment specification than the one proposed by this definition. Usually, a factorization in the form of

$$E[YW|\lambda_1, \lambda_2] = E[E(Y|W; \lambda_1)W|\lambda_1, \lambda_2],$$

(1.1)

is sufficient to detect the causal impact of $W$ on $Y$. The problem is that, even for simple economic situations, it is often hard to justify an assumption like (1.1).

Consider for example the case where an economist wants to study the demand function of soft drinks using the individual consumption of Coca-Cola ($X$), the individual consumption of Pepsi-Cola ($Q$) and their respective prices ($p_1, p_2$) (with $p_1 > p_2$). A typical dataset looks like the one in Figure 1.1.
Figure 1.1: $X^*$ is the consumption of Coca, while $Q^*$ is the consumption of Pepsi.

From the observation of the data, an econometrician would conjecture that, while the first two cross-section observations (i.e. individuals) may consider Coca-Cola and Pepsi-Cola as perfect substitutes, and therefore, since $p_1 > p_2$, all the income spent on soft drinks goes to Pepsi-Cola, the individuals 3 and 4 prefer to consume a quantity of Coca-Cola $X^*$ different from zero, even though the price of Coca-Cola is higher (see Figure 1.2).

In other words, since Agent 1 and Agent 3 have different preferences, their optimization process is different

\[
\begin{align*}
\text{Agent 1 max process} & : & \text{Agent 3 max process} \\
\max_{X,Q} U(X,Q) = \beta_1 X + \beta_2 Q & \quad \text{s.t.} I = p_1 X + p_2 Q \leq I_{ub} \\
\max_{X,Q} U(X,Q) = X^\beta_1 Q^{\beta_2} & \quad \text{s.t.} I = p_1 X + p_2 Q \leq I_{ub}
\end{align*}
\]

with $I_{ub}$ being the budget constraint. In order to check whether the previous conjecture is true or not, a structural model that enables us to empirically validate the hypothesized choice structure must be specified. If the utility functions are not cardinal, the results of the two maximization processes cannot be compared directly. To the contrary, the study of the expenditure functions allows to monetize the otherwise incommensurable trade-offs between the benefits of the consumptions and their costs. In particular, an expenditure function indicates the minimum amount of money that an individual would need to spend in order to achieve
Figure 1.2: The (individual) demand functions for a given budget constraint $p_1 X + p_2 Q = I$ varies accordingly to individual preferences. Agent 3 and 4 do not consider Coca-Cola and Pepsi to be equally good (left graph), while Agent 1 and 2 do (right graph).

A certain level of utility (given an utility function and a set of prices). If the conjectured choice models are correct, then for those agents that consider Coca-Cola and Pepsi-Cola as perfect substitutes (like Agent 1), the expenditure function should be

$$I(p_1, p_2, \bar{v}(X, Q)) = \min \left( p_1 \frac{\beta_1 X + \beta_2 Q}{\beta_1}, p_2 \frac{\beta_1 X + \beta_2 Q}{\beta_2} \right),$$

where $\bar{v}(X, Q)$ is the level of utility for the observed consumption levels $(0, Q)$. For individuals that do not consider the two soft drinks as perfect substitutes (like Agent 3), the amount of expenditures for the given level $(X, Q)$ should be

$$I(p_1, p_2, \bar{v}(X, Q)) = \left[ p_1 \left( \frac{\beta_1 P_1}{\beta_2 P_2} \right)^{\frac{\beta_2}{\beta_1 + \beta_2}} + p_2 \left( \frac{\beta_1 P_1}{\beta_2 P_2} \right)^{\frac{\beta_1}{\beta_1 + \beta_2}} \right] X^{\frac{\beta_1}{\beta_1 + \beta_2}} Q^{\frac{\beta_2}{\beta_1 + \beta_2}}.$$

In this case both Definition 1 and/or assumption (1.1) are useless, because the required factorization for the vector $W = [X, Q]^T$, given a set of parameters $\lambda_1 = \cdots$
[\beta_1, \beta_2]$, can be true for the perfect substitute case or for the imperfect one, but not for both.

This simple introductory example shows that when micro-data exhibit holes (non-participation in the activity of interest), kinks (switching behaviors) and corners (non-consumption or non-participation at specific points in time), then relations like (1.1) become meaningless (Pudney, 1989). There are at least three solutions to deal with an assumption like (1.1) in a context where heterogeneity among individuals is a major concern.

A first solution is to aggregate the data and study a much smoother problem (being smoother due to the compensations of the movements in opposite directions) typical for macro-data. Consider, for example, a relation between two variables which at a micro level may be piecewise linear with many nodes. After the aggregation is done, the relationship can probably be approximated by a smooth function that can satisfy equation (1.1) (Eilers and Marx, 1996). However, if an econometrician is interested in the analysis of individual-level data, in order to describe the economic behavior of individuals or firms, this option does not help.

A second possibility is to accept the heterogeneous nature of the parameters at a micro-level, but to ignore it, and use a parametric (possibly linear) specification with constant coefficients. Let us now abstract from the above example and denote the response by $Y$ and the two explanatory variables $X$ and $Q$ such that

$$Y_i = t(X_i, Q_i) + e_i = \beta_0 + \beta_1 X_i + \beta_2 Q_i + \epsilon_i \quad E[\epsilon_i | X_i, Q_i] = 0 . \quad (1.2)$$

In this case all the heterogeneity is absorbed by the regression disturbance $\epsilon$. Even if many applied economists recognize the limits of a standard parametric specification that most likely suffers from a functional form misspecification because $t(X_i, Q_i) \neq \beta_0 + \beta_1 X_i + \beta_2 Q_i$, which means that $e \neq \epsilon$, they still use it as an approximation because their least squares estimates (like OLS) converge to the value of $\beta$ that tries to minimize (we say try because its success depends also
on other factors like the scedasticity function) the mean-squared prediction error
\( E[t(X_i, Q_i) - \beta_0 - \beta_1 X_i - \beta_2 Q_i]^2 \). It is well known that under homoscedasticity OLS gives the best linear predictor of the non-linear regression function (the mean-squared error (MSE) being the loss function), and this even when there is a functional form misspecification (White, 1980b). However, this property is not useful if the objective of the researcher is to interpret the regression coefficients as a true micro-relationship in the form of \( E[Y_i|X_i, Q_i] \), because the standard OLS would typically be inconsistent when estimating the marginal effect of the variables,

\[
\hat{\beta}_1^{OLS} \to \beta_1 \neq \frac{\partial t(X_i, Q_i)}{\partial X_i} =: \beta_{1i} \quad \hat{\beta}_2^{OLS} \to \beta_2 \neq \frac{\partial t(X_i, Q_i)}{\partial Q_i} =: \beta_{2i}.
\]

In particular, if the returns are heterogeneous in the data generating process (DGP), a modeling strategy like (1.2) might not be able to derive consistent estimates. For example, if \( Y_i = \beta_i^T W_i + \varepsilon_i \) with \( W = [1, X, Q]^T \), for \( \beta_i = [\beta_{0i}, \beta_{1i}, \beta_{2i}] \) is modeled as

\[
Y_i = \beta^T W_i + \varepsilon_i \quad \text{with} \quad \beta = E[\beta_i] \quad \text{and} \quad \varepsilon_i = W_i^T [\beta_i - \beta] + \varepsilon_i,
\]

then the standard OLS estimators would give

\[
\hat{\beta}_{OLS} = \left[ \sum_{i=1}^{n} W_i W_i^T \right]^{-1} \sum_{i=1}^{n} W_i Y_i = \left[ \sum_{i=1}^{n} W_i W_i^T \right]^{-1} \sum_{i=1}^{n} W_i \left[ W_i^T E(\beta_i) + \varepsilon_i \right]
\]

\[
= \frac{p}{n \to \infty} E(\beta_i) + E(W_i W_i^{-1} E(W_i W_i^T [\beta_i - E(\beta_i)]) + E(W_i W_i^T) E(W_i \varepsilon_i)
\]

\[
= E(W_i W_i^T)^{-1} E(W_i W_i^T \beta_i).
\]

From the last equality it follows that \( \hat{\beta}_{OLS} \to E[\beta_i|W_i] \) unless \( E[\beta_i|W_i] = E[\beta_i] \).

A third solution is to conclude that the discreteness and non-linearity typical for micro-data requires to model heterogeneity directly. But how? A first option is to transform the density requirement of Definition 1 into an individual level factorization like

\[
f(y_i, w_i|\lambda_{1i}, \lambda_{2i}) = f(y_i|w_i; \lambda_{1i}) f(w_i|\lambda_{2i}) \quad i = 1, \ldots, n \quad (1.3)
\]
(here \(w_i\) needs not to include a 1 for the intercept), where every cross-sectional observation is characterized by a set of individual parameters \((\lambda_{1i}, \lambda_{2i})\). This creates the complication that the parameters \((\lambda_1, \lambda_2)\) are no longer variation free, which is not \textit{stricto sensu} a problem because it is possible to transform \(\lambda_{1i}\) into a random coefficient to which it is possible to associate an invariant hyper-parameter \(\theta\) that characterizes the prior density \(f(\lambda_{1i}|w_i, \theta)\). In this specification, the invariance assumption can be reproduced in the form \(f(y_i|w_i, g(w_i, \theta))\), where \(\theta\) is estimated globally by a maximum likelihood or in a neighborhood, e.g. by a kernel-based local likelihood. This Bayesian solution allows to have variation-free hyper-parameters and, at the same time, random coefficients that capture individual heterogeneity due to the randomness of \(\lambda_{1i}\).

No matter how elegant the solution might look like, it presents many and interdependent problems. The main one is the low degree of robustness of the estimates \(\hat{\theta}\). One may use shrinking priors to overcome this, but in order to make sure that the prior decays quickly enough (to produce robust estimates), it is necessary to impose stringent conditions both on the priors’ tails and on the decay rates of the tails. This kind of assumptions are very hard to understand in practice and even harder to relate to economic theory.

A less controversial way to directly model heterogeneity is to allow the value of the coefficients to change when and observable variable \(F_i\), called here ‘effect modifier(s)’, allows to write equation (1.2) as

\[
Y_i = \beta_i^T W_i + \varepsilon_i \quad \text{with} \quad \beta_i = g(F_i) + \delta_i .
\]

(1.4)

This is the well-known varying coefficient model (VCM), cf. Hastie and Tibshirani (1993). In this specification \(Y_i\) is the dependent variable, \(W_i\) is a \(d_W \times 1\) vector of explanatory variables, and the coefficient \(\beta_i\) is allowed to vary across \(i\). In particular, it is a function of a \(d_F \times 1\) vector of observable variables \(F_i\) (which might include also elements of \(W\)), while \(g(.)\) is a vector of functions of the effect
modifier, and $\delta_i$ is a stochastic mean-zero disturbance with finite variance. The exogeneity assumption is centered on the idea of correctly estimating the causal impact of $W$, not the one of $F$, on $Y$, therefore it is possible to imagine $\hat{g}$ as the best nonparametric predictor of $\beta_i$ for a given $F_i$. This implicates that the expected value of $\delta$ given $Q$ would be equal to zero by construction: $E[\delta_i|F_i] = 0$. The new structure of the model produces a very flexible and yet interpretable semiparametric specification.

The hybrid nature of the VCMs has several advantages. Firstly, it reduces the level of complexity of a pure nonparametric model allowing to interpret the coefficients like in a parametric specification. Secondly, it enables to incorporate insights that come from economic theory into the modeling process. Thirdly, it produces a good trade-off between the loss in fitting ability, which is (hopefully) small compared to the nonparametric specification, and the increased facility of the estimation process, which is almost as easy as in a parametric model.

The empirical potentials of the VCM modeling can be understood re-considering the soft drink example. In this case, depending on whether the agent considers the goods as perfect substitutes or not, the coefficients resulting from the optimal allocations are different. However, in both cases, they are functions of the level of expenditure $I$, the prices $(p_1, p_2)$ and the quantity consumed $(X, Q)$ by individuals with some characteristics also included in $F$.

The previous consideration suggests that a VCM, in which the returns are functions of the prices and of the quantities of the goods, allows to keep a linear specification for the expenditure function (or expenditure shares) in the form of

$$Y_i = \beta_{0i} + \beta_{1i}X_i + \beta_{2i}Q_i + \varepsilon_i \quad (1.5)$$

with $\beta_j = g_j(F_i) + \delta_{ji}$, $j = 0, 1, 2$. In other words, a VCM allows us to transform the structural specification of (1.2) into a model able to take into account heterogeneity sive natura, making an assumption like (1.1) meaningful and often also plausible.
Of course, the presence of numerous effect modifiers makes an equation like (1.5) hard to compute. To the contrary, a function with few effect modifiers is more easily interpretable and, at the same time, reduces the curse of dimensionality of the nonparametric regression \( \beta_j = g_j(F_i) + \delta_{ji}, j = 0, 1, 2 \). Therefore it makes sense to reduce the number of effect modifiers for each \( j \) separately (e.g. by canceling those that present a low level of non-linearity with respect to the regressors).

The introduction of a second, more complex, economic example helps explaining the potentials of a VCM, even when the conjectures about the individual-decision making process behind the observed covariates is less easy to deduce than in a simple demand analysis environment. Let’s suppose that an applied economist wants to study the impact of education and experiences on wages in a cross-sectional data set. The concerns about the disaggregated nature of the data might induce the researcher to do an a priori analysis that most likely reveals that marginal returns to education vary for different levels of working experience, see e.g. Schultz (2003). Merging the insights that come from the economic theory with the intuitions resulting from the scrutiny of the data we end up with a VCM of the form

\[
\text{wage}_i = \beta_0 + \beta_1 \text{educ}_i + \varepsilon_i, \tag{1.6}
\]

where the intercept and the slope are functions of the level of experience, \( \beta_i = g(\text{exp}_i) + \delta_i \), with \( g(\cdot) \) and \( \delta \) belonging to \( \mathbb{R}^2 \). A structural specification like (1.6) is very appealing because it corrects the (downward) bias that would emerge using a linear modeling that ignores the interaction between experience and education and therefore systematically underestimates the returns on schooling (Card, 2001). In this new formulation, it is important to discuss the role of \( \delta \). As indicated above, the nature of \( \delta \) is not the one of an isotonic deviation from the mean but rather the one of a stochastic disturbance in a nonparametric equation. Therefore the role that \( \delta \) plays in equation (1.4) is related to its disturbance-nature.

Unlike in the soft drinks example, where the choices’ structure was easy to reverse-
Reverse engineering, also called back engineering, is the process of extracting knowledge or design information from anything man-made, and re-producing it. In economics, the reverse engineering process consists of extracting the structure of individual preferences from observed outcomes and then reproduce the outcomes using the conjectured informations.
here heterogeneity can explicitly imply deviations from the slopes. In the semi-parametric framework, the quantity of local deviation $\delta$ is a function of the degree of smoothness of $g(exp)$. At the same time, since the primary objective of the research is not to estimate correctly the causal impact of $F$ on $Y$ but rather the one of $W$, it is sufficient to think of $\hat{g}(\cdot)$ as the best nonparametric predictor of $\beta_i$ such that $E[\delta_i|F_i] = 0$ becomes true by construction. As a result, the average returns to education are equal for all the cross-section observations that have the same level of $exp$.

1.2 Triangular Varying Coefficient Models with Instruments

The previous section highlighted the necessity to model heterogeneity directly in order to make the assumptions of Definition 1 plausible. But still, exogeneity can be violated by the nature of the observed variables irrespectively of the semiparametric characteristics of the VCM. In particular, a regressors could be endogenous in the sense that in the structural equation $Y_i = \beta_i^T W_i + \varepsilon_i$ one has $E[\varepsilon_i|W_i, F_i] \neq 0$. The three usual sources of endogeneity typically mentioned are: the omission of explanatory variables correlated with the included covariates, a measurement error and reversed causality. All the three sources of endogeneity cannot be solved using the varying coefficient approach alone.\footnote{However, the most typical, though in economics rarely mentioned, endogeneity problem, i.e. the functional misspecification, can be largely diminished by the VCM.} A popular solution is to introduce some additional variables called instruments.

Definition 2. A variable $Z$ is called an instrumental variable (IV) for $W$ if

1. is partially correlated with the endogenous variable $W$ once the other explanatory variables have been netted out.
2. is mean-independent with respect to the stochastic error $\varepsilon$.

This definition suggests that the addition of a second structural equation to the VCM creates a triangular model able to exogenize $W$ while modeling heterogeneity directly. For simplification, let us set for a moment $\dim(W) = \dim(Z) = 1$. Keeping a specification like (1.4), it is sufficient to add a selection equation that relates the endogenous $W$ with the instrument(s) $Z$, namely

$$W_i = m(Z_i) + \eta_i \quad E[\eta_i|Z_i] = 0 \quad (1.7)$$

and assume a finite variance for $\eta_i$. In this formulation the vector of explanatory variables is allowed to contain endogenous components, while $Z$ is a vector of IVs, which may have $F$ as one of its arguments. Furthermore, $m(\cdot)$ is a smooth function, or a vector of smooth functions if $\dim(W) > 1$, while $\varepsilon$ and $\eta$ are, respectively, the endogenous error and a stochastic disturbance that has expected value equal to zero and finite variance.

The triangular nature of equations (1.4) and (1.7) implies a simple endogeneity mechanism. In order for the error term $\varepsilon$ to be correlated with at least one of the explanatory variables $W$, it must be $\text{cov}(\eta, \varepsilon) \neq 0$. To see how the mechanism of the model works in practice, it is useful to consider the simplest possible specification, namely a model that would include only one heterogeneous intercept, one heterogeneous slope and one endogenous explanatory variable. The latter is instrumented by one exogenous variable $Z_1$, which is correlated with $W$ even if the impact of the (exogenous) effect modifier has been netted out, namely

$$Y_i = \beta_0 + \beta_1 W_i + \varepsilon_i \quad E[\varepsilon_i|F_i, W_i] \neq 0$$
$$W_i = m(F_i, Z_{1i}) + \eta_i \quad E[\eta_i|F_i, Z_{1i}] = 0 \quad .$$

In this specification, irrespectively of the relation between the error $\varepsilon$ and the two disturbances $\delta$ and $\eta$, endogeneity comes only through $\text{cov}(\varepsilon, \eta)$, see causal graph 1.4.
Figure 1.4: The mechanism of the endogeneity process changes depending on the assumptions about the relationship between the error $\varepsilon$ and the stochastic disturbances $(\eta, \delta)$. The left picture is the only possibility in a world of homogeneous coefficients, while the right specification (with $cov(\eta, \delta) \neq 0$) is the situation resulting from introducing a varying coefficient structure. The direct connection between $\delta$ and $\varepsilon$ is not taken into account because the interest is about the causal link between $Y$ and $W$ for a given level of $F$.

The considerations about the mechanisms of the endogeneity problem combined with the observation that a VCM is a special case of a semiparametric linear specification, suggest that the model can be identified and later estimated using the control function approach (Tesler, 1964). The control function, say $h(\cdot)$, handles the relation between $\eta$ and $\varepsilon$ (irrespectively of the behavior of $\delta$) in the following form

$$
\epsilon_i = \delta_i W_i + \varepsilon_i = h(\eta_i) + \theta_i \quad E[\theta_i | \eta_i] = 0 .
$$

(1.8)

This added to (1.4) eliminates the endogeneity problem giving unbiased estimates for $g(\cdot)$:

$$
E[Y_i | Z_i, \eta_i] = g(F_i) W_i + h(\eta_i) \quad Z_i = (F_i, Z_{1i}) .
$$

(1.9)

It is important to notice that the higher complexity of a VCM increases the chance to successfully eliminate the endogeneity problem via the control function approach. Specifically, even if a set of valid instruments $(Z_i)_{i=1}^n$ is available,
a linear IV estimator would generally be biased. For example, if the equation
\[ Y_i = \beta_0 + \beta_1 W_i + \epsilon_i \] (with \( \epsilon_i \perp Z_i \)) is modeled using homogeneous coefficients
\[ Y_i = \beta_0 + \beta_1 W_i + e_i \] with \( e_i = [\beta_0 - \beta_0] + W_i[\beta_1 - \beta_1] + \epsilon_i \) and \( \beta_j = E(\beta_j) \),
\( j = 0, 1 \), then the instrumentation using \( Z_i \) does not produce consistent estimates.
Consider for example the case where \( \text{dim}(Z) = \text{dim}(W) \geq 1 \). In this setting the estimated returns are
\[
\hat{\beta}_i^{IV} = \left[ \sum_{i=1}^{n} Z_i W_i^T \right]^{-1} \sum_{i=1}^{n} Z_i Y_i = \left[ \sum_{i=1}^{n} Z_i W_i^T \right]^{-1} \sum_{i=1}^{n} Z_i \left[ W_i^T E(\beta_i) + e_i \right]
\]
\[
\xrightarrow{p} E(\beta_i) + E(Z_i W_i^T)^{-1} E(Z_i W_i^T [\beta_i - E(\beta_i)]) + E(Z_i W_i^T)^{-1} E(Z_i \epsilon_i)
\]
\[
= E(Z_i W_i^T)^{-1} E(Z_i W_i^T \beta_i).
\]
The last equality cannot be simplified further unless a new assumption, namely \( \beta_i \perp (W_i, Z_i) \), is made - which is clearly in contradiction with the spirit of the model, cf. the causal graphs in Figure 1.4. Basically, the heterogeneous nature of the returns transforms \( Z_1 \) into a ‘poor’ instrument if the simple linear structure is used.

In order to proceed and correctly estimate the unknown terms in equation (1.9), it is necessary to impose additional identification conditions. Identification can be obtained imposing a conditional mean independence in the form of
\[
E[\epsilon_i | Z_i, \eta_i] = E[\epsilon_i | \eta_i] \quad \text{CMI}, \tag{1.10}
\]
or a conditional moment restriction
\[
E[\epsilon_i | Z_i] = 0 \quad \text{CMR}. \tag{1.11}
\]
The CMI and the CMR are not equivalent (Kim and Petrin, 2013). The CMI requires \( Z \) and \( \eta \) to be additively separable in \( W \), which often is not the case. To the contrary, the CMR can be easily justified by the use of economic primitives that describe the structural specification (Benini and Sperlich, 2016). The use of
the CMR, however, requires to include the instrument(s) in the control function, such that the relation between $\epsilon$ and $\eta$ becomes

$$
\epsilon_i = h(Z_i, \eta_i) + \vartheta_i \quad E[\vartheta_i | Z_i, \eta_i] = 0 \, .
$$

(1.12)

In any case, if the amplitude of the control function increases, a less precise estimate $\hat{g}(.)$ might be produced (multi-functionality). This is the statistical counterpart of the econometric problem called ‘weak’ instruments, i.e. instruments that are weakly correlated with the endogenous regressors.

The estimation of VCM in its simplest specification has been proposed in different forms. Hastie and Tibshirani (1993) used a smoothing spline based on a penalized least squares minimization, while Fan and Zhang (2008) proposed a kernel weighted polynomials. However, this last method and its surrogates are designed for a single effect modifier for all coefficients, which is a strong limitation in the context we discussed so far.

Estimating an equation like (1.9) is a more complicated procedure than the one required for a simple VCM. The presence of a control function, which depends upon $\eta$, requires the use of specific tools that are designed for additive models. The two most common alternatives are the marginal integration method (Linton and Nielsen, 1995) and the smooth back-fitting (Roca-Pardinas and Sperlich, 2010). The latter method suffers less from the curse of dimensionality and can be applied as part of a 2-steps procedure. The first step consists in the estimation of $m(Z)$ in equation (1.7) using a standard nonparametric technique. The second step consists in the substitution of the estimated residuals $\hat{\eta}$ into (1.9), which creates an equation characterized by a finite sample disturbance whose impact can be mitigated asymptotically (Sperlich, 2009). For an exhaustive survey on the VCM estimation techniques see Park, Mammen, Lee and Lee (2013). For a comparison of implementations of these methods in R (R Core Team, 2014), including the control function approach, see Sperlich and Theler (2015).
All the previous considerations are particularly important in the treatment effect literature. For a discrete $W$ with finite support, Imbens and Angrist (1994,1995) named the impact of a treatment (i.e. a change in $W$) local average treatment effect (LATE). By construction, the LATE can only compute the average of the $\beta_i$ for the individuals that choose to switch their $w$ because of an instrument’s change. In other words, in the LATE environment, the parameter of interest can only be estimated for people responding to the selection equation and is therefore an instrument (or selection) specific parameter. They imposed a conditional independence assumption in the form of $Y_i(w) \perp Z_i \forall w$, as well as the request of independence of the (so called) compliers sub-population to an instrument’s change. Reconsidering these assumptions in the presence of heterogeneous returns shows that the LATE is not defined if $\text{cov}(\beta, Z) \neq 0$. In the case of a VCM this means that, unless the effect modifier is indeed a constant, the standard independence assumption used to define and identify the LATE is not fulfilled.

The model we outlined above suggests that, if some effect modifiers $F_i$ are observed, they should be used to construct a VCM that makes the LATE conditions more credible. For example, in the case of a binary endogenous treatment $W$ which is instrumented by a binary instrument $Z$, the varying LATE becomes

$$LATE(q) = \frac{E[Y | F = q, Z = 1] - E[Y | F = q, Z = 0]}{E[W | F = q, Z = 1] - E[W | F = q, Z = 0]}.$$ 

Integrating over $q$ gives the value of the LATE. In this case, the more heterogeneity of returns to $W$ is captured by $g(F)$ the less the LATE will vary over the IVs’ choice. In other words, a VCM reduces the typical LATE problem to a minimum because it controls for the correlation between the effect modifier and the instrument. Therefore, the VCM enables to identify a LATE-type parameter that can be estimated nonparametrically regressing $Y$ and $W$ on $F$ and $Z$. The interesting point here is that the parameter of interest depends on both, the instruments’ choice and the values taken by $F$. An interesting next step would be to find a meaningful model specification that merges the effect modifier and the
1.3 An Example

In order to see all the potentials of the triangular VCM specification in practice, it is useful to reconsider the wages-experience-education relationship. Experience and education are crucial variables in the determination of a worker’s wage. Yet, labor economists have argued for many years that cognitive and non-cognitive abilities are also critical in order to determine labor market outcomes. A large empirical literature has confirmed the positive connection between cognitive test scores and high wages (Murnane, Willett and Levy, 1995). Unfortunately, many datasets do not provide ability’s measures. The lack of information about the skills misleads the researcher to mistake the data generating process (DGP). Even if a VCM modeling strategy is used, if the ability of the individual is not included, such that

\[ wage_i = t(educ_i, exp_i, ability_i) + \zeta_i \]

is modeled as

\[ wage_i = g_0(\exp_i) + g_1(\exp_i)educ_i + \epsilon_i, \]

then the exogeneity assumption \( E[\epsilon_i|educ_i, exp_i] = 0 \) does not hold, because of an omitted variable bias. This problem can be solved using an instrument.

There exist at least two classical errors that arise when searching for an IV. The first one is the selection of a variable that is clearly correlated with the endogenous regressor but hardly independent from the error \( \epsilon \). For example, the level of education (of one) of the parents would be hardly independent from the omitted variable \( ability \). A second wrong choice would be the selection of a variable that has the opposite characteristics, namely a variable that is exogenous but that is hardly correlated with the endogenous regressor. For example, the last digit of the person social security number. The choice of good instruments must come from both, a deep knowledge of the origin of the IV and the source of endogeneity.
Take instead the example proposed by Angrist and Krueger (1991). In most American states education legislation requires students to enter school in the calendar year when they turn six. Therefore the age at which students start school is a function of the date of birth of the pupils. For example, if the 31st of December is the legal cut-off point, children born in the fourth quarter enter school shortly before turning six, while those born in the first quarter enter school when they are around six years and an half. Furthermore, because compulsory schooling laws require students to remain in school only till they turn 16, these groups of students will be in different grades, or through a given grade to a different degree, when they reach the legal drop out age. The combination of the school start-age policies and the school attendance laws creates a situation where children attend school for different times depending upon their birthdays. Assuming that the day of birth of a person is not correlated with his abilities seems to make the quarter of birth \((qob)\) a valid IV. The typical mistake made here is to conclude from no-causality to no-correlation. But firstly, there is clearly the possibility that the IV is correlated with the education of the parents, and secondly, being the youngest could mean to be the smallest and physically weakest in the class resulting in maturity disadvantages. All these facts could change the wage-path invalidating the IV.

Nonetheless, let us consider a VC triangular model with the same instrument proposed by Angrist and Krueger

\begin{align}
\text{wage}_i &= g_0(exp_i) + g_1(exp_i)\text{educ}_i + h(\eta_i) + \vartheta_i \\
\text{educ}_i &= m(exp_i, qob_i) + \eta_i.
\end{align}

(1.13) \hspace{1cm} (1.14)

In this specification they identify the LATE of education on wages for those who do not drop out in spite of the ones that could have thanks to their birth date. Note that, if this is not the parameter of interest, it might have been much better and easier to use a proxy approach instead of an IV one. In order to reverse-engineer the preferences’ structure it is necessary to model a situation where a rational
individual has to decide, when turning 16, to stay for the rest of the academic year or leave school. In this context, the agent’s wage is a function of the years of education \( educ \), but also of his unobserved ability, \( \varepsilon \). The agent’s ability is not observed, but the information set that the student can consult before the decision to stay or not is made includes a signal of his individual ability \( \eta \), for example his past grades. The cost to stay until the end of the year is a function of an exogenous cost shifter, namely the quoter of birth \( qob \), if a student turns 16 in January the cost to stay till the end of the year is higher than if he turns 16 in May, so it makes sense to consider the quoter of birth an argument of the cost function. At the same time, the agent’s utility has to be function of the education’s choice, the cost-shifters and the unobserved ability, \( U(educ, qob, \varepsilon) = p(educ, \varepsilon) - c(educ, qob) \), where \( p(\cdot) \) is the education production function and \( c(\cdot) \) is the cost function. The optimal choice problem becomes

\[
\text{educ} = \arg\max_{educ} \{ E[U(\text{educ}, qob, \varepsilon)|qob, \eta] \}. \tag{1.15}
\]

The specification of the utility function is crucial. The functional form \( U(educ, qob, \varepsilon) = p(educ, \varepsilon) - c(educ, qob) \) is not chosen for convenience. The quarter of birth must be part of the cost function, otherwise \( qob \) would not be valid instruments – but at the same time it cannot be part of the educational production function because otherwise the causal effect of \( educ \) cannot be excluded from the joint effect of \( (educ, qob) \). The costs can depend among the ability’s signal, \( \eta \), if for example a staying-based financial aid is available. This possibility, however, is not taken into account. The decision problem just described is illustrated in Figure 1.5.

In this context the exclusion restriction requires the choice variable \( educ \) to be separable in \( (\varepsilon, qob) \). This depends upon the assumptions that the researcher is willing to make about the educational production function \( p(\cdot) \) and the cost function \( c(\cdot) \).

All the previous considerations show how a model like (1.13)-(1.14) is able to: 1.
Figure 1.5: When endogeneity is an issue due to the presence of a model misspecification, the use of VCM is not enough to guarantee causal analysis, and the introduction of IVs becomes necessary to ensure the exogeneity.

- make individual returns heterogeneous,
- solve the endogeneity problems that are due to the functional form misspecification using the VCM nature of the model,
- solve the endogeneity problems that are due to the nature of the regressors using IVs, and
- relate the structural specification to the economic theory providing a rigorous microfoundation of the outcome equation.
Bibliography


Chapter 2

Modeling Heterogeneity: A Praise for Varying-coefficient Models in Causal Analysis*

2.1 Introduction

In economic theory, agents aim to achieve their highest level of welfare depending on the return of their choices. This return is heterogeneous in the sense that it varies over agents even after having conditioned on all kind of observable covariates. Heterogeneous returns have always been an issue of modeling beyond the simple adding of additive independent errors (and apart from modeling these as heteroscedastic, autocorrelated, spatial dependent, etc.). This is one of the reasons why the varying-coefficient models (Hastie and Tibshirani, 1993, or Fan and Zhang, 1999), the mixed- or random-effects models (see González Manteiga, Lombardía, Martínez Miranda, and Sperlich, 2013, for a recent review), and the Bayesian modeling (see for example Kneib and Fahrmeir, 2006) have been attracting an

*Published with Stefan Sperlich at Computational Statistics, volume 30, pages 693-718.
increasing attention. This holds especially true for the area of non- and semiparametric statistics from which all these reference are taken. Like the well-known additive nonparametric model (Hastie and Tibshirani, 1990, Mammen, Linton and Nielsen, 1999) they have been proposed in order to bridge the gap between nonparametric models and the standard (generalized) linear model. They have been designed to help the non- and semiparametric methods to get from theory to practice. Especially in empirical economics, non- and semiparametric methods have so far only been used for some prediction issues in finance. Nowadays, they are also proposed for matching in the treatment effect literature. But even there they do rather exist in the theory than in practice. However, there is an increasing awareness of problems that happen to arise with the identification, estimation and interpretation of average returns or effects. If heterogeneity in returns is present, then OLS estimation in linear models is inconsistent, and instrumental variable (IV) estimators may lose their identification capacity: they need to satisfy quite unrealistic and non-testable conditions, but for less strong assumptions identify only the average effects of non-identified subpopulations (the so-called LATE = local average treatment effect), if at all. Consequently, impact evaluation becomes much harder - no matter whether ex-post (inference) or ex-ante (prediction).

Micro foundations of economic theory are often concerned about mathematical modeling of agents under strong rational choice and utility optimization. When facing the problem of generalizing the individual results to the whole population, the existence of a mean or representative agent behavior is used to be assumed. While already this assumption has been a controversial issue in economic theory since half a decade (Allais, 1953), the related econometric problems have usually been underestimated. More precisely, as deviations from the mean behavior are typically endogenous, the simple traditional inference leads to biased and inconsistent estimates. A classical and well studied example is the resulting downward bias of the consumer price index, see for example Diewert (1996) or Crossly and
In the present article, we start with considering the case of the linear regression model, i.e. the benchmark in econometric modeling:

\[ Y_i = X_i^T \beta + \varepsilon_i \quad i = 1, ..., n \]

where \( \beta \) is supposed to be the vector of coefficients quantifying the return (or say, effect) of vector \( X_i \) on scalar \( Y_i \). As a consequence, \( \varepsilon_i \) is an error term with mean 0, being mean-independent of \( X \) such that there is 'a priori' no endogeneity problem. Heterogeneity comes from \( X \) and the error term, and the latter must capture all individual deviation from the mean return. Not only from a micro-econometric point of view, this is a quite restricted heterogeneity which does not match with the modern individual oriented modeling. For example, by considering the impact of the years of schooling on the wage, it seems obvious that the return of schooling varies between individuals due the personal abilities. If these are not included in \( X \), one speaks of an omitted variable problem, but if there is an interaction with education, it is also a modeling or functional form related problem. Consequently, a more realistic but (if applied without further modeling) too general model is

\[ Y_i = X_i^T \beta_i + \varepsilon_i \quad i = 1, ..., n , \quad (2.1) \]

where the heterogeneity is also reflected in \( \beta_i \). If one is only interested in the average return or effect, \( E[\beta_i] = b \), the following model is actually the one of interest:

\[ Y_i = X_i^T b + e_i , \quad e_i = X_i^T [\beta_i - b] + \varepsilon_i , \quad i = 1, ..., n . \quad (2.2) \]

Heterogeneity in the returns does not invalidate the OLS estimation as long as the deviation from their mean is exogenous, i.e. \( E[\beta_i - b|X_i] = 0 \). However, this condition does typically not hold in economics as this would to some extent contradict the spirit of rational behavior of individuals. When agents chose their actions, they consider the personal abilities that impact the return of their choices.
Students choose the number of years of schooling by taking into account their personal abilities which will affect their future wage in many ways. Therefore, the assumption of exogeneity of \( X_i \) is violated in this model in the sense that \( E[e_i|X_i] \neq 0 \) and the OLS estimation becomes inconsistent.

A popular way to deal with endogeneity in econometrics is to use a linear IV (or, equivalently, a two step least squares 2SLS) estimation. This requires to be provided with an additional variable \( W_i \) such that \( \text{Cov}(W_i, e_i) = 0 \) but \( \text{Cov}(W_i, X_i) \neq 0 \) for \( i = 1, ..., n \). Nevertheless, in practice this is almost never the case. Indeed, to find a set of observable variables \( W_i \) that strongly impacts \( X_i \) but with \( E[e_i|W_i] = 0 \) is extremely unlikely. Already the existence of \( W_i \) with \( \text{Cov}(W_i, \beta_i) = 0 \) is hard to justify unless \( W_i \) is completely randomly drawn (like a lottery number) and that has no other relation to \( \beta_i \) than that via \( X_i \). Similar arguments apply if we interpret the heterogeneity in returns as an effect of omitted variables. As an example take Card (1995) who assumed that the proximity to a college has an effect on the years of schooling but no further effect on the wage. At a first glance, it seems to be a valid instrument. However, the proximity to school is linked with the social background which is partly determined by the individual abilities and thus, the return of schooling. This includes that all people working at the school live close to it, and at the same time care more about the education of their children than the average of parents does.

To sum up, in presence of those 'endogenous' coefficients \( \beta_i \), the traditional approaches of OLS and IV estimation fail in their attempt to estimate an average effect. Indeed, it is difficult to obtain an average effect without estimating directly the heterogeneity (nonparametrically). In a different setup, similar conclusions have been drawn by Deaton and Muellbauer (1980) in their studies on consumer behavior and consumer price index estimation.

An alternative way to deal with the dependence between \( \beta_i \) and \( X_i \) is to consider
the use of a factor that controls for (at least most of) this dependency, say a third variable $W_i$. If such a $W_i$ is available with $E[\beta_i|W_i]$ being the best prediction of $\beta_i$ given $W_i$, then $\beta_i - E[\beta_i|W_i] = u_i$ with $E[u_i|W_i] = 0$ by definition. With $g(\cdot)$ being this best predictor, equation (2.2) becomes

$$Y_i = X_i^T E[\beta_i|W_i] + X_i^T u_i + \varepsilon_i = X_i^T g(W_i) + X_i^T u_i + \varepsilon_i, \quad i = 1, \ldots, n. \quad (2.3)$$

The variables comprised in $W$ are often called effect modifiers. This provides a natural extension of the above mentioned varying-coefficient model

$$Y_i = X_i^T g(W_i) + \tilde{\varepsilon}_i, \quad E(\tilde{\varepsilon}_i|X_i, W_i) = 0 \quad i = 1, \ldots, n. \quad (2.4)$$

For $W_i$ related to $X_i$ one hopes to control by this means for the dependency between $X_i$ and $\beta_i$. In other words, to estimate heterogeneity, the coefficient can be modeled as a nonparametric function (i.e. an unknown but deterministic smooth function) of an observed variable $W_i$ up to a random deviation $u_i$ that is mean-independent from $X_i$. The effect modifier helps us to mitigate or solve the endogeneity that may occur due to heterogeneous returns or effects. Note that variable $W$ must not be independent from $X$ if it shall filter potential dependency between $\beta_i$ and $X_i$. An extreme case is setting $W = X$ giving basically the well-known nonparametric additive model.

Effective estimation of non- and semiparametric varying-coefficient models has been introduced in various articles, see Hastie and Tibshirani (1993), Fan and Zhang (1999, 2000, 2008), including recursive estimation to improve efficiency (Cai, Fan, and Li, 2000), estimation under measurement errors (Chiang, Rice and Wu, 2001), models with generated regressors (Pendakur and Sperlich, 2010), generalized varying-coefficient models (Mammen and Nielsen, 2003, Roca-Pardiñas and Sperlich, 2010) or additive varying coefficients (Yang, Park, Xue, and Härdle, 2006). The well-known time varying-coefficient models for longitudinal data should be mentioned as well. For a recent review see Mammen, Park, Lee and Lee (2013). The extension of (2.4) to (2.3) is just an extension of the classical nonparametric
varying-coefficient model toward a non- or semiparametric mixed effects model, see for example González-Manteiga, Lombardía, Martínez-Miranda and Sperlich (2013).

In econometrics, however, these models are hardly known. The potential of applying this type of model has not been sufficiently emphasized. The aim of this article is to highlight the advantage of varying-coefficient models to deal with (endogenous) heterogeneity, and to study the state-of-the-art of implementation and calculation. On this purpose, econometric justification as well as a discussion of computational means in the statistic software R (see R Core Team, 2014) will be provided.

The rest of this article is structured as following: In the next section the problems that occur when using classical linear methods (including IV estimators) are illustrated. In Section 2.3 we show why and how varying-coefficient models can mitigate these problems. Section 2.4 briefly describes the implementation of these methods and the available R packages. These are then compared by simulation studies. Section 2.5 concludes.

2.2 Problems with linear OLS and IV estimation

Identifying an average effect or return is a multidimensional problem. When we interpret the coefficient in a linear model, we say \( Y \) would increase or decrease ‘in average’ by the value of the coefficient if one changed \( X \) by one unit. This is often misunderstood as being the average return (to \( X \)) in general. This is unfortunately incorrect as it refers only to the averaging-out of the isotonic error iff this is mean-independent from \( X \). When the heterogeneity of returns is stochastically related to \( X \) (causality is not required), this interpretation is definitively wrong. Let us attempt therefore to clarify this a bit.
Intuitively, to study an overall average effect, $E[\beta_i]$ makes sense if we can obtain the following equality from equation (2.2)

$$E[Y_i|X_i] = X_i^T E[\beta_i], \quad i = 1, ..., n.$$  

This equation is satisfied if $E[e_i|X_i] = 0$, i.e. the orthogonality due to the exogeneity assumption of $X_i$ with respect to $\epsilon_i$, but also $E[\beta_i|X_i] = E[\beta_i] = b$. The latter is the mean-independence of $\beta_i$ from returns $X_i$. Violation of this independence leads to

**Proposition 1.** *The OLS Estimator of the average effect $E[\beta_i]$ in model (2.2) is generally, i.e. without further assumptions, inconsistent.*

**Proof.** Let $\{(Y_i, X_i)\}_{i=1}^n$ be a sample of independent random variables drawn model (2.2) with $E[\epsilon_i|X_i] = 0$. Then,

$$\hat{\beta}_{OLS} = (\Sigma X_i X_i^T)^{-1} \Sigma X_i Y_i \xrightarrow{p} E[\beta_i] + E^{-1}[X_i X_i^T] E[X_i \epsilon_i]$$

$$= E[\beta_i] + E^{-1}[X_i X_i^T] E[X_i X_i^T (\beta_i - E[\beta_i]) + X_i \epsilon_i]$$

$$= E^{-1}[X_i X_i^T] E[X_i X_i^T \beta_i].$$

From the last line we will see that $\hat{\beta}_{OLS} \xrightarrow{p} E[\beta_i]$ unless $E[\beta_i|X_i] = E[\beta_i]$ which is not true in general.

Under the following simulations design, we get an idea of the consequences and
the dimension (i.e. the severity) of this result:

\[ Y_i = 0.6 + X_i \beta_i + \varepsilon_i \]

\[ X_i = \gamma W_i + u_i \]

\[ \beta_i \sim Beta(1.65 W_i, 5) \]

\[ \varepsilon_i \sim N(0, Var(X_i \beta_i)) \]

\[ W_i \sim \log N(0, 1) \]

\[ u_i \sim N(0, 1) \]

\[ \gamma = \frac{Corr(X_i, W_i)}{\sqrt{Var(W_i)} - Corr(X_i, W_i)^2 Var(W_i)} . \]

We computed the average estimated bias of the OLS estimator \( \frac{1}{m} \sum_{j=1}^{m} \hat{\beta}_{j, OLS} - E[\beta_i] \) from \( m = 1000 \) simulation samples, each of size 100. Table 2.1 gives the biases for intercept (0.6) and the average slope (0.2857). As expected, the bias increases with the correlation between the coefficients and the regressors although there is not necessarily a (direct) causality relation from \( X \) on \( \beta \). We see clearly that albeit a nontrivial dependency structure between \( \beta_i \) and \( X_i \), the 'bias' of an OLS estimate compared to the average slope is quite severe.\(^1\)

The result, that the OLS estimator is biased, is neither new nor surprising. The reason why this small simulation is presented here is to get aware of the role of \( g, W_i \) and \( u_i \) when switching to model (2.3). This enables us to better understand source and size of the bias, to potentially model it, and to discuss ways of circumventing it – those that work and those that do not.

But is this simulated design of interest, does it reflect somehow any realistic situation, in particular in economics? We think, it does. Take for example economies of scale – like they are generally accepted for the agricultural sector. The main interest is to estimate and understand the production function. 'Economies of scale'\(^1\)

\(^1\)We put bias in quotation marks as the OLS estimate is actually not biased in the statistical sense, it does simply estimate something different than the average slope. Alternatively you may say that not the estimator is wrong but the interpretation that people often attach to it.
Table 2.1: Simulated Bias over the Correlation between W and X

<table>
<thead>
<tr>
<th>Corr($x_i, w_i$)</th>
<th>Intercept</th>
<th>Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.95</td>
<td>0.5169</td>
<td>0.4061</td>
</tr>
<tr>
<td>-0.75</td>
<td>0.0510</td>
<td>0.2426</td>
</tr>
<tr>
<td>-0.5</td>
<td>-0.0328</td>
<td>0.1140</td>
</tr>
<tr>
<td>-0.25</td>
<td>-0.0315</td>
<td>0.0315</td>
</tr>
<tr>
<td>0</td>
<td>-0.0024</td>
<td>0.0017</td>
</tr>
<tr>
<td>0.25</td>
<td>0.0288</td>
<td>0.0287</td>
</tr>
<tr>
<td>0.5</td>
<td>0.0359</td>
<td>0.1117</td>
</tr>
<tr>
<td>0.75</td>
<td>-0.0554</td>
<td>0.2413</td>
</tr>
<tr>
<td>0.95</td>
<td>-0.5198</td>
<td>0.4085</td>
</tr>
</tbody>
</table>

| Parameter Expected | 0.6 | 0.2857 |

means that the size of an entity (enterprise, farm, ...) changes the expected returns. Severance-Lossin and Sperlich (1999) studied the additive marginal effects and elasticities of inputs for farms in Wisconsin, USA. They applied a standard nonparametric additive model of the form

$$\log Y_i = \gamma_0 + \sum_{j=1}^q \gamma_j (\log X_{i,j}) + \varepsilon_i$$ (2.5)

with all inputs $X$ and output $Y$ measured in US-Dollar, where all $\gamma_j$ functions were nonparametric corresponding therefore to our model when setting $\gamma_j (\log X_{i,j}) = g_j (\log X_{i,j}) \cdot \log X_{i,j}$ or $\gamma_j (\log X_{i,j}) = g_j (X_{i,j}) \cdot X_{i,j}$. They indeed found increasing returns $\sum_j \gamma_j' > 1$ for middle-sized and large farms; this means, size made the $\gamma_j$ functions non-linear such that $X_i$ would be endogenous in a standard log-linear model. Size was clearly driving both (at least partly), inputs $X_i$ and $\beta_i$, and are therefore our effect modifiers $W$.

Similarly, in order to study the efficiency of labor offices in the Czech Republic during the transition years, Profit and Sperlich (2004) estimated a dynamic fixed
effects panel model with time fixed effects $\eta_t$ and individual fixed effects $\delta_i$

$$
Y_{i,t} = \gamma_0 + \sum_{j=1}^{4} \gamma_j (\log X_{i,t-1,j}) + \sum_{1 \leq j < k \leq 4} \gamma_{j,k} (\log X_{i,t-1,j}, \log X_{i,t-1,k}) + \delta_i + \eta_t + \epsilon_{i,t}
$$

(2.6)

which is basically an extension of the additive interaction model of Sperlich, Tjøstheim and Yang (2002) for longitudinal data. Equation (2.6) is a standard – apart from the fact that it is nonparametric – economic model to study job-matching with $Y$ being log unemployment to job exits, and $X$ summarizing the lagged $Y$, the log number of vacancies, the log number of inflows to unemployment, and the unemployment stock. Profit and Sperlich (2004) tested the returns to scale of the labor offices, studied time and space variation (i.e. the non-uniformity of Job-Matching), and afterward the sources of their variation. They found that the variation (i.e. the non-linearity of the $\gamma_j$ functions) was due to borderland, and specific active labor market policies, say a vector of effect modifiers $W$. At the same time, this vector $W$ was clearly correlated to $X$ with different causality directions: e.g. being borderland has a positive impact on unemployment stock, but unemployment stock has an impact on the active labor market policy.

For both examples one may argue that a varying-coefficient models had been reasonable alternatives which allowed to include $W$ directly in the model. Subsequently, a further step would be to consider $\beta_i - g(W_i)$ as an unobservable disturbance that then may lead to a similar (though now mitigated) endogeneity problem. One could argue that $W$ might be hard to find or too weak to control for the dependency between $X$ and the returns. Therefore, in econometrics, a popular way to deal with all kind of endogeneity is the IV estimation. This approach is based on a set of additional variables, so-called instruments $Z_i$, that on the one hand is supposed to be exogenous in the sense that $E[e_i|Z_i] = 0$, but on the other hand correlated with $X_i$. As discussed above, this is extremely unlikely to exist.

In general, such a moment condition does not hold, leading to
Proposition 2. For a random sample \( \{ Z_i \}_{i=1}^n \) of 'instruments' with \( E[\varepsilon_i | Z_i] = 0 \), \( \text{Cov}(X_i, Z_i) \neq 0 \), the linear IV estimator of \( b = E[\beta_i] \) in model (2.2) is generally inconsistent except if \( E[\varepsilon_i | Z_i] = 0 \), \( E[\beta_i - b | Z_i] = 0 \) is not sufficient.

Proof. For simplification but without loss of generality let \( \dim(X) = \dim(Z) \).

Then

\[
\hat{\beta}_{IV} = (\Sigma Z_i X_i^T)^{-1} \Sigma Z_i Y_i = (\Sigma Z_i X_i^T)^{-1} \Sigma Z_i \{ X_i^T E[\beta_i] + \varepsilon_i \}
\]

\[
\xrightarrow{p}{n \to \infty} E[\beta_i] + E^{-1}[Z_i X_i^T] E [Z_i X_i^T \{ \beta_i - E[\beta_i] \}] + E^{-1}[Z_i X_i^T] E[Z_i \varepsilon_i]
\]

\[
= E^{-1}[Z_i X_i^T]^{-1} E[Z_i X_i^T \beta_i],
\]

where the last equality cannot be simplified without further assumptions like for example \( \beta_i \perp (X_i, Z_i) \) for all \( i = 1, ..., n \).

As said, the independence of \( \beta_i \) with \( X_i \) and \( Z_i \) is unlikely and even contradicts economic theory as discussed above. However, assuming \( \beta_i \perp Z_i \) is not enough because this does typically not give \( E[Z_i X_i^T (\beta_i - b)] = 0 \), needed for the consistency. But for this latter case Imbens and Angrist (1994, 1996) defined the local average treatment effect (LATE). More specifically, under a set of additional non-testable assumptions, the linear IV estimator identifies the average return (or effect) for 'compliers'. These form the subpopulation of individuals that increased their value of \( X \) due to an increase in \( Z \). It is indispensable that all individuals move in the same direction (monotonicity assumption); therefore we can not relax the term 'increase' to simply 'change'. Its derivation relies on the identification of the expected potential outcome for a given value of the instrument \( z \) with the conditional expectation: \( E[X_i^T (Z_i) \beta_i] = E[Y_i | Z_i = z] \). Clearly, is satisfied because \( \beta_i \perp Z_i \) staying the crucial independence assumption for LATE. But as the compliers form only a specific subpopulation defined by the choice of instruments \( Z \), the LATE is not necessarily a relevant parameter. Finally one should mention that the concept was introduced for binary \( X \) and \( Z \). Extensions are tedious but straight for discrete
$X$ and $Z$ with finite support. The new LATE becomes than a weighted average of the LATEs of compliers for stepwise increasing $Z$.

To sum up, an estimator based on a LATE argument would inherit the same defects as the classic IV estimator although it seems to allow to relax the independence assumption. Its interpretation is only clear for binary $X$ and $Z$, but even then may refer to an irrelevant parameter. In any case it will give only an average which in turn is little informative if heterogeneity in returns is important. The inadequacy of the tradition inference motivates the use of a new class of estimators. The proposal in the next section is based on a non-parametric modeling of the coefficient by considering the use of a set of observed variables $W_i$, often called 'effect modifiers'.

### 2.3 The varying-coefficient model for heterogeneous effects

Let us now consider the varying-coefficient model (2.3). Introducing effects modifiers $W_i$ helps to mitigate the harmfulness of the association between $\beta_i$ and $X_i$, and therefore $\beta_i$ and $Z_i$. One may even re-establish the validity of $Z_i$ as an instrument 'conditionally on $g(W_i)$'. At this point, a crucial assumption is required for the present section: $u_i \perp X_i$ or $u_i \perp Z_i$ for all $i = 1, ..., n$. In other words, $\beta_i$ depends on $X_i$ or $Z_i$ only through $W_i$.

For brevity we discuss simply the two situations where (1) $u_i$ is mean-independent from $X_i$, and (2) when (1) fails but either $\beta_i(w)$ or $u_i$ are mean-independent from an instrument $Z_i$ which fulfills $E[\varepsilon|Z_i] = 0$ with $E[X_i|Z_i]$ not being a constant. Other cases like independence between $X_i$ and $W_i$ are not of interest here.
2.3.1 When $u$ is mean-independent from $X$ given $W$

Recall that equation (2.3) was just a different way of writing $Y_i = X_i^T \beta_i + \varepsilon_i$ but splitting $\beta_i$ into a part conditioned on $W_i$ and a mean-zero remainder term $u_i$ such that

$$E[Y_i|X_i] = X_i^T E[g(W_i)|X_i] \neq X_i^T E[g(W_i)]$$

unless $X_i$ and $W_i$ are mean-independent. In this case, it is hardly possible to compute an average effect without conditioning on $W$, too. In contrast, doing so you get

$$E[Y_i|X_i, W_i] = X_i^T g(W_i) ,$$

or simply

$$Y_i = X_i^T g(W_i) + e_i ,$$

with $E[e_i|X_i, W_i] = 0$. This is the simple varying-coefficient model (2.4) which can be estimated by any known semiparametric method.

It should be reminded that standard IV methods will fail here. This example illustrates the fact that if we want to obtain an overall parameter estimation, it is hardly avoidable to estimate the observable dependent heterogeneity in a first step. This is particularly true for cases where heterogeneity may be of first order. This is were varying-coefficient models are good for. Let us give two examples from the literature.

Making the Engle curves more flexible has been a central issue when analyzing consumer behavior in microeconomics. To our knowledge, in this context semiparametric varying-coefficient models have been applied first by Pendakur and Sperlich (2010). They studied consumer behavior, and modeled the expenditure shares by varying-coefficients models. More specifically, $Y_i$ is now a vector of expenditure shares for $M$ goods,

$$Y_i = \text{exp.shares}_i = \gamma(E_i^R) + \Gamma(E_i^R) P_i ,$$

(2.7)
where vector $\gamma(\cdot) \in \mathbb{R}^M$ and matrix $\Gamma(\cdot) \in \mathbb{R}^{M \times M}$ are nonparametric functions of the real total expenditures $E^R_i$ of household $i$ (which are predicted from the model).\(^2\) $P_i$ is the vector of the log prices of the $M$ goods faced by household $i$. This allows one to e.g. check for Slutsky symmetry or to determine the effective inflation index for any income group. Asymmetry and different effective inflation indices imply that returns to prices must be heterogeneous, and this is what the authors found.

Similarly, Pendakur, Scholz, Sperlich (2010) considered the indirect utility in consumer demand systems. They modeled the utility for household $i$ as

$$U_i = E_i - P^T_i \gamma(E_i) - \frac{1}{2} P^T_i \Gamma P_i,$$

where $E_i$ it reported total expenditure of household $i$, $P_i$ again the vector of log prices of all $M$ items, $\Gamma$ an unknown $M \times M$ matrix of unknown coefficients, and $\gamma$ again a vector of $M$ nonparametric functions. Roy’s identity (Roy, 1947) tells us that the vector of expenditure shares is given by $\frac{\partial U}{\partial P} / \frac{\partial U}{\partial E}$, i.e. in our case

$$\text{exp.shares}_i = \frac{\gamma(E_i) + \Gamma P_i}{1 - P^T_i \gamma'(E_i)}.$$  \hfill (2.9)

Equation (2.9) can be estimated from observed data, although not directly with standard methods. Again, the authors found heterogeneous returns to prices in the utility (2.8) of households. No direct method could have directly estimated the average returns to prices.

\subsection*{2.3.2 Dependence between $u$ and $X$ given $W$ but existence of an instrument $Z$}

As indicated above, we consider the situation where $Z_i$ is mean-independent from $u_i$ once conditioned on $W_i$. For this case a LATE type parameter can be identified.

\(^2\)The preposition real indicates that the numbers $E^R$ assigned to the different households - in this example from all-over Canada - refer to the same purchase power.
This is based on the joint conditioning of \( W_i \) and \( Z_i \) combined with the traditional monotonicity assumption related to this kind of inference. On this purpose and for the sake of simplicity, \( X_i \) is considered to be a binary random variable. For simplicity, also \( Z_i \) is considered as a binary random variable. The extension to discrete variables with finite support is straight but rather tedious in notation and interpretation. One can then establish

**Definition 3.** Let \( \{(Y_i, X_i, W_i)\}_{i=1}^n \) be a random sample generated by model (2.3), and \( \{Z_i\}_{i=1}^n \) with \( E[\varepsilon_i|W_i, Z_i] = E[u_i|W_i, Z_i] = 0 \) such that \( X_i \) and \( Z_i \) are dependent binary random variables with \( X_i(Z_i = 1) - X_i(Z_i = 0) \geq 0 \) almost sure for all \( i \).

The conditional LATE evaluated at \( Z_i = z \) is

\[
\beta_{LATE}(w) = \frac{E[Y_i|W_i = w, Z_i = 1] - E[Y_i|W_i = w, Z_i = 0]}{E[X_i|W_i = w, Z_i = 1] - E[X_i|W_i = w, Z_i = 0]}.
\]

This is the expected average effect of \( \beta_i(w) = g(w) + u_i \) in the population of compliers for the given value \( w \) of the effect modifier \( W_i \). If there are no compliers for \( W_i = w \), i.e. \( E[X_i|W_i = w, Z_i = 1] - E[X_i|W_i = w, Z_i = 0] = 0 \), then \( \beta_{LATE}(w) \) is not defined. We can estimate this conditional LATE now simply by combining the classical Wald estimator for endogenous dummies with a nonparametric estimator for varying-coefficient models.

Note, however, that also this LATE type estimator identifies only an average effect on the subpopulation of compliers which do not constitute a clearly identified (sub)population or might not be of interest. So how does this improve compared to the classic LATE? It does in the following two aspects. First, depending on the choice and strength of the effect modifier \( W \), it is much easier to argue that \( Z_i \) was independent from \( u_i \) than arguing it was independent from \( \beta_i \). Second, if \( W \) is a strong modifier, we can control and thus study for most of the heterogeneity; the different parameters identified by the different instruments are expected to have a small variation, i.e. the choice of instruments should mainly effect the variance but hardly the mean of the estimator. A generalization to \( X_i, W_i, Z_i \).
being continuous is therefore appealing. This can be done by the so-called control function approach, see Telser (1964). Newey, Powell and Vella (1999) introduced a nonparametric extension and made it popular in econometric theory.

Let us consider the following simultaneous equations system:

\[ Y_i = X_i^T g(W_i) + X_i^T u_i + \varepsilon_i \]  
\[ X_i = h(Z_i) + v_i, \quad i = 1, ..., n, \]  

with the same assumptions as above but now \( X \) and \( Z \) being (potentially) continuous and \( E[v_i|Z_i] = 0 \). This is a natural extension of model (2.3). Note that the additivity in equation (2.11) guarantees the monotonicity of the effect of \( Z \) on \( X \) (assumed in the binary case). Clearly, \( h \) can be nonparametrically estimated in order to predict \( v_i \) by \( \hat{v}_i := X_i - \hat{h}(Z_i) \). Then,

\[ E[Y_i|X_i, Z_i, v_i] = X_i^T g(Z_i) + X_i^T E[u_i|X_i, Z_i, v_i]. \]  

As all dependence between \( X_i \) and \( u_i \) is reflected in \( v_i \), and \( Z_i \) was independent from \( u_i \) by definition, we get for \( E[u_i|X_i, W_i, v_i] = E[u_i|v_i] =: \lambda(v_i) \) now

\[ E[Y_i|X_i, W_i, v_i] = X_i^T g(W_i) + X_i^T \lambda(v_i) \]
\[ = X_i^T g(W_i) + X_i^T \lambda(\hat{v}_i) + X_i^T [\lambda(v_i) - \lambda(\hat{v}_i)], \]

where from Sperlich (2009) we know that the last term is negligible. Moreover, it is known from the literature on additive varying coefficients that \( g_i \) and \( \lambda_i \) are identifiable up to additive and multiplicative constant terms.

### 2.4 Estimating varying-coefficient models in practice

We already presented above a list of references introducing different estimators for various versions of varying-coefficient models. In this section we provide a summary
of existing implementations for varying-coefficient model estimators. Namely, we
discuss some important R (R Core Team, 2014) packages that contain those meth-
ods. However, this review is not pretended to be complete or exhaustive. It rather
serves as an illustration that practitioners are already today provided with the
necessary tools for performing the modeling and estimation strategies discussed in
the preceding section.

Among others, Fan and Zhang (1999, 2000, 2008) proposed a smoothing method
for estimating $g$ in a classical varying-coefficient model $Y_i = X_i^T g(W_i) + \varepsilon_i$ with
$E[\varepsilon_i|X_i, W_i] = 0$ for all $i = 1, \ldots, n$. Their method was based on kernel weighted
local polynomials. Intuitively it is quite appealing because it exhibits strong sim-
ilarities with a weighted least squares M-estimator but only at a local level. In
the local linear case one considers $g(W) \approx \alpha_0 + \alpha_1^T (W - w_0)$ in a neighborhood
of $w_0$ with $\alpha_0 \in \mathbb{R}^{\text{dim}(X)}$ and $\alpha_1 \in \mathbb{R}^{\text{dim}(W) \times \text{dim}(X)}$. More specifically, one tries to
minimize the squared error

$$E \left[ (Y - \{X^T \alpha_0 + X^T \alpha_1^T (W - w_0)\})^2 | W = w_0 \right].$$

Given a sample $\{(Y_i, X_i, W_i)\}_{i=1}^n$ of independent variable from the same model,
by convolution with a kernel function $K_h(\cdot)$ provided with bandwidth $h$ around a
point of interest $W_0$, one obtains the empirical counterpart:

$$(a_0^*, a_1^*) = \arg\min_{(\alpha_0, \alpha_1)} \frac{1}{n} \sum_i \{Y_i - [X_i^T \alpha + X_i^T b(W_i - z_0)]\}^2 K_h(W_i - z_0),$$

where $a_0^*$ can be shown to be a consistent estimator for $g(z_0)$. Differently from
this kernel smoothing approach, Hastie and Tibshirani (1993) proposed a smooth-
ing spline method, also based on a penalized least squares minimization. There,
each coefficient is parametrized on a cubic spline basis. Local maximum likelihood
(e.g. Cai, Fan and Li, 2000), other spline methods (e.g. Chiang, Rice and Wu,
2001), smooth backfitting (Mammen and Nielsen, 2003; Roca-Pardinas and Sper-
lich, 2010) and Bayesian structured additive models (Fahrmeir Kneib and Lang,
have later on be introduced as well. For most of these methods asymptotic theory has been provided. Inference is typically based on resampling methods or on rough (parametric plug-in) approximations. But even more interesting for the empirical researcher is the question regarding implementation. Nowadays, several R packages can be used to estimate varying-coefficient models.

Hayfield and Racine (2008) have developed the so-called (np) package (nonparametric kernel smoothing methods for mixed data types). This package enables us to implement a kernel estimation for a wide range of variable types: continuous variables, ordered or unordered factors. The function npscoef provides estimators for a varying-coefficient model. An optimal bandwidth $h$ can be obtained by a least-squares cross-validation method applying the function npscoefbw. A problem of this package is that it seems hard (if not impossible) to modify the set of effect modifiers from one varying coefficient to another, i.e. all varying coefficients depend on the same set $W$. It is also not able to estimate additive varying coefficients. Note that the function predict only returns predictions for the expected $Y$. It does not predict the estimated coefficient functions at arbitrary values of the effect modifier. This makes a fair comparison with other packages a bit difficult.

Wood (2011, 2006, 2004, 2003, 2000) proposed a computationally fast algorithm for estimating generalized additive models based on penalized regression splines. This is implemented in his package mgcv. The available distribution families can be found with the function family.mgcv. The function considered in the simulation study below is gam.

Not exactly the same but quite related, at least in many situations, is the gamlss package. It was developed by Rigby and Stasinopoulos (2005), and originally devoted to the estimation of generalized additive models. The common exponential families are replaced by a more flexible distribution family allowing for extreme kurtotic, skewed or over-dispersed continuous and discrete distributions. The available
distribution families can be found with the function "gamlss.family". The special case of varying-coefficient models is handled by a maximum likelihood method based on penalized B-splines. The function studied in the following is gamlss().

All the considered packages have been designed to implement the estimation of semi-parametric additive models from which the varying-coefficient model is a particular case. As a result, the implementation of the latter does not require any specific adaptation of the used packages. After using the function gam() or gamlss(), the predict() function with the option terms enables to extract each additive term evaluated at the wanted points. Then, it is sufficient to divide the term by the related regressor to obtain the varying-coefficient function predictions. The implementation of the control function approach, considered in the last simulation design above, is a multi-stage estimation that enables us to implement a specific regression at every single step of the estimation. Details are provided below.

In order to compare the three packages, several simulation designs are used. Due to the variety of methods used by these packages, the designs share common features to enable a comparison on the same basis: (a) the coefficient functions are the same whatever the design is; (b) a uniform distribution of the effect modifiers to prevent a bias at the tails when using a fixed bandwidth in kernel estimation, i.e. make kernel and spline methods comparable; (c) bounded coefficient functions with trigonometric specification to stabilize the variance; (d) predictions of the coefficients functions at equidistant points on the interval $[-2, 2]$.

We simulated four different designs, see below. From each we draw 1500 independent random samples for sample sizes $n = 50, 100, 250$ (i.e. 500 simulation runs for each size) to obtain 500 replicates at fixed values of the effect modifiers $W$ for each design and sample size. From them we calculated 90% simulation intervals (i.e. cut the outer 5% from above, and from below), see Figures 2.1 to 2.12, and the average bias (ABias) and average MSE (AMSE) for the varying coefficient functions.
In all figures the dotted lines in the center are the 'true' data generating function, the circles are the averages of the 500 predictions, and the bands are constructed from the 5 and 95% quantiles of the 500 simulations. Note that for notational convenience the two nonparametric coefficient functions in Design 1 to 3 will be denoted by $g(\cdot), h(\cdot)$ instead of $g_1(\cdot), g_2(\cdot)$.

### Design 1

As previously mentioned, prediction of the coefficient isn’t implemented in the np package. Further, this package does not seem to be designed to consider different sets of effect modifiers for each coefficient function. However, a preliminary simulation study allows to appreciate the performance of this package in case these problems are irrelevant for the practitioner. This is the aim of the first design where the effect modifier is considered as being a fixed grid on the interval $[-2, 2]$.

Else, we simulated the model

$$Y = g(W) + h(W)X + \varepsilon$$

$$g(W) = \cos(3W) , h(W) = \sin(2.5W)$$

$$X = 2W_1 - 0.5W_2^2 + 2W , W = \left\{-2 + \frac{4i}{n} | i = 1, \ldots, n \right\}$$

$$W_1 \sim N(0, 1) , W_2 \sim N(0, 1) , \varepsilon \sim N(0, 1) .$$

The results are given in Figures 2.1, 2.2 and 2.3, and Table 2.2. The figures lead us to the conclusion that the np and mgcv perform pretty well. When looking The mgcv package seems to dominate other packages in terms of ABias and AMSE when the sample size is small ($n = 50$). For larger sample sizes, the np package performs much better. Indeed, when the curvature is important ($g(\cdot)$), there is a slight bias that is probably due to the smoothing effect related to any nonparametric estimation that the np package tends to reduce as the sample size increases. The
Figure 2.1: Design 1, 90% simulation intervals for sample size $n = 50$ for $g$ on the left, $h$ on the right.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Package</th>
<th>$ABias[g(.)]$</th>
<th>$ABias[h(.)]$</th>
<th>$AMSE[g(.)]$</th>
<th>$AMSE[h(.)]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>mgcv</td>
<td>0.0262</td>
<td>0.0035</td>
<td>0.0641</td>
<td>0.0088</td>
</tr>
<tr>
<td></td>
<td>np</td>
<td>0.0518</td>
<td>0.0073</td>
<td>0.037</td>
<td>0.0179</td>
</tr>
<tr>
<td></td>
<td>gamlss</td>
<td>0.0262</td>
<td>0.2088</td>
<td>0.2348</td>
<td>4.0957</td>
</tr>
<tr>
<td>100</td>
<td>mgcv</td>
<td>0.0364</td>
<td>-0.0021</td>
<td>0.0267</td>
<td>0.0032</td>
</tr>
<tr>
<td></td>
<td>np</td>
<td>0.0251</td>
<td>0.0003</td>
<td>0.0168</td>
<td>0.0077</td>
</tr>
<tr>
<td></td>
<td>gamlss</td>
<td>0.0364</td>
<td>-2.2790</td>
<td>0.1460</td>
<td>517.4362</td>
</tr>
<tr>
<td>250</td>
<td>mgcv</td>
<td>0.0425</td>
<td>-0.0014</td>
<td>0.0129</td>
<td>0.0014</td>
</tr>
<tr>
<td></td>
<td>np</td>
<td>0.015</td>
<td>0.0005</td>
<td>0.0073</td>
<td>0.0034</td>
</tr>
<tr>
<td></td>
<td>gamlss</td>
<td>0.0425</td>
<td>-0.6461</td>
<td>0.1335</td>
<td>1348.804</td>
</tr>
</tbody>
</table>

Table 2.2: Results calculated from 500 samples drawn from Design 1.

The `gamlss` package estimation has clearly a problem for $W \leq 0$ when estimating the 'intercept' function $g(.)$, and is rather imprecise at values $W \in [0, 1]$ when
Figure 2.2: Design 1, 90% simulation intervals for sample size $n = 100$ for $g$ on the left, $h$ on the right.

estimating $h(.)$. We could not find out what exactly is the source of the problem. This imprecision tends to even increase when the sample size increases.

**Design 2**

The second design is related to case where the effect modifiers are simply the regressors of the model, i.e. $X = W$. Again the functions $g, h$ are univariate but now refer to different effect modifiers. For this reason $np$ is no longer included. This, however, allowed us to now consider random variables $W$ that are independently drawn in each simulation study, as both, *gamlss* and *mgcv* are able to predict $g, h$ on fixed grid no matter what the original effect modifier was. The simulated design
Figure 2.3: Design 1, 90% simulation intervals for sample size $n = 250$ for $g$ on the left, $h$ on the right.

was

\[ Y = 6 + g(X_1)X_1 + h(X_2)X_2 + \varepsilon \]

\[ g(X_1) = \cos(3X_1), \quad h(X_2) = \sin(2.5X_2) \]

\[ X_1 \sim U[-2, 2], \quad X_2 = 0.25X_1 + 0.75U[-2, 2], \quad \varepsilon \sim N(0, 1). \]

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Package</th>
<th>$ABias[\hat{g}(\cdot)]$</th>
<th>$ABias[\hat{h}(\cdot)]$</th>
<th>$AMSE[\hat{g}(\cdot)]$</th>
<th>$AMSE[\hat{h}(\cdot)]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>mgcv</td>
<td>-0.2152</td>
<td>0.0259</td>
<td>0.2172</td>
<td>0.1053</td>
</tr>
<tr>
<td></td>
<td>gamlss</td>
<td>-0.0004</td>
<td>0.0003</td>
<td>0.5283</td>
<td>2.3782</td>
</tr>
<tr>
<td>100</td>
<td>mgcv</td>
<td>0.1393</td>
<td>0.0106</td>
<td>0.0932</td>
<td>0.0597</td>
</tr>
<tr>
<td></td>
<td>gamlss</td>
<td>0.043</td>
<td>0.0012</td>
<td>0.5169</td>
<td>3.2682</td>
</tr>
<tr>
<td>250</td>
<td>mgcv</td>
<td>-0.0831</td>
<td>0.0038</td>
<td>0.0366</td>
<td>0.0233</td>
</tr>
<tr>
<td></td>
<td>gamlss</td>
<td>0.0647</td>
<td>-0.0043</td>
<td>0.8001</td>
<td>5.3895</td>
</tr>
</tbody>
</table>

Table 2.3: Results calculated from samples drawn from Design 2.
The results are plotted in Figures 2.4, 2.5 and 2.6, and in Table 2.3. The mgcv package manages to estimate the coefficient functions but with a bias where the curvature is important. However, this problem decreases with the sample size. So this might simply be a smoothing effect. This is different in the gamlss package that just fails to fit the coefficient function. A direct analysis of ABias and AMSE tends to underestimate these two issues because negative and positive bias compensate each other all along the support of the coefficient functions. In any case it is obvious that this design should better be estimated by (generalized) additive models as the implementation in the implementation of the varying-coefficient models in the gamlss package seems to require different (or maybe even independent?) variations for the regressors $X$ and the effect modifiers $W$. We say already in design 1 that this package has problems where the relation between $X$ and $W$ was strong.
Design 3

The design is no crossing the dependence structure, i.e. the effect modifier $W_1$ for the coefficient of $X_1$ is correlated with $X_2$, and $W_2$ with $1$. Else one can speak of a standard bivariate nonparametric varying-coefficient model so that there should be no particular identification problem for \texttt{gamlss}. The simulated design in detail is

$$Y = 6 + g(W_1)X_1 + h(W_2)X_2 + \varepsilon$$

$$g(W_1) = \cos(3W_1), \ h(W_2) = \sin(2.5W_2)$$

$$W_1 \sim U[-2, 2], \ W_2 = 0.25W_1 + 0.75U[-2, 2]$$

$$X_1 = 0.25W_1 + 0.4W_2 + W_2, \ X_2 = 2W_1 - 0.5W_2^2 + 2W_1$$

$$W_1 \sim N(0, 1), \ W_2 \sim N(0, 1), \ \varepsilon \sim N(0, 1).$$

The numerical results are shown in Figures 2.7, 2.8 and 2.9, and Table 2.4. Both
Figure 2.6: Design 2, 90% simulation intervals for sample size $n = 250$ for $g$ on the left, $h$ on the right.

Table 2.4: Results calculated from 500 samples drawn from Design 3.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Package</th>
<th>$ABias[\hat{g}(.)]$</th>
<th>$ABias[\hat{h}(.)]$</th>
<th>$AMSE[\hat{g}(.)]$</th>
<th>$AMSE[\hat{h}(.)]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>mgcv</td>
<td>-0.0043</td>
<td>0.0040</td>
<td>0.0230</td>
<td>0.0193</td>
</tr>
<tr>
<td></td>
<td>gamlss</td>
<td>0.0312</td>
<td>-0.0720</td>
<td>11.2338</td>
<td>2.9439</td>
</tr>
<tr>
<td>100</td>
<td>mgcv</td>
<td>-0.0310</td>
<td>-0.0166</td>
<td>0.0190</td>
<td>0.0124</td>
</tr>
<tr>
<td></td>
<td>gamlss</td>
<td>0.2775</td>
<td>-0.0574</td>
<td>24.0567</td>
<td>0.4506</td>
</tr>
<tr>
<td>250</td>
<td>mgcv</td>
<td>-0.0007</td>
<td>-0.0022</td>
<td>0.0040</td>
<td>0.0066</td>
</tr>
<tr>
<td></td>
<td>gamlss</td>
<td>0.1755</td>
<td>0.2812</td>
<td>4.6757</td>
<td>5.2134</td>
</tr>
</tbody>
</table>

packages, mgcv and mgcv manage to predict the shape of the coefficient functions. However, the mgcv clearly performs better. Similar conclusion can be drawn from Table 2.4. Note that the identification issue (design 2) has disappeared, and the smoothing effect mentioned in design 1 seems to be reduced here.
Figure 2.7: Design 3, 90% simulation intervals for sample size \( n = 50 \) for \( g \) on the left, \( h \) on the right.

**Design 4**

This design studies the control function approach introduced in Section 2.3.2. Note that in this design function \( h(\cdot) \) is no longer a standard varying coefficient but plays the role of the control function. The simulated design is now

\[
Y = 6 + g(W)X + uX + \varepsilon
\]

\[
g(W) = \cos(3W), \quad h(v) = \sin(2.5v)
\]

\[
W \sim U[-2, 2], \quad X = 0.25W + 0.4Z_2 + v
\]

\[
Z_1 \sim N(0, 1), \quad Z_2 \sim N(0, 1)
\]

\[
u = h(v) + N(0, 1), \quad v \sim U[-2, 2], \quad \varepsilon \sim N(0, 1).
\]

At the first step, \( v \) can be predicted by regressing \( X \) on \( W \) and \( Z_2 \) with any non-parametric method. At the second step, \( h(\cdot) \) is estimated on \( \hat{v} \) jointly with \( g(\cdot) \) on \( X \). For a fair comparison we did this first step for all simulations and packages.
Figure 2.8: Design 3, 90% simulation intervals for sample size \( n = 100 \) for \( g \) on the left, \( h \) on the right.

applying \texttt{gam} no matter whether the second step was then performed with \texttt{mgcv} or \texttt{gamlss}. So if only one of the method fails, then this is due to the second step; if both fail it might be due to the first and / or second step.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Package</th>
<th>( ABias[\hat{g}(.)] )</th>
<th>( ABias[\hat{h}(.)] )</th>
<th>( AMSE[\hat{g}(.)] )</th>
<th>( AMSE[\hat{h}(.)] )</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>mgcv</td>
<td>0.0262</td>
<td>0 (-9.222657e-17)</td>
<td>0.0390</td>
<td>0.3151473</td>
</tr>
<tr>
<td></td>
<td>gamlss</td>
<td>0.0233</td>
<td>0.4281</td>
<td>2.4170</td>
<td>6.0226</td>
</tr>
<tr>
<td>100</td>
<td>mgcv</td>
<td>0.0364</td>
<td>0 (8.939166e-17)</td>
<td>0.0128</td>
<td>0.1210</td>
</tr>
<tr>
<td></td>
<td>gamlss</td>
<td>-1.0389</td>
<td>-3.1684</td>
<td>105.6606</td>
<td>1072.2730</td>
</tr>
<tr>
<td>250</td>
<td>mgcv</td>
<td>0.0425</td>
<td>0 (-9.794682e-17)</td>
<td>0.0056</td>
<td>0.0321</td>
</tr>
<tr>
<td></td>
<td>gamlss</td>
<td>-0.0598</td>
<td>-0.9288</td>
<td>1.6459</td>
<td>178.423</td>
</tr>
</tbody>
</table>

Table 2.5: Results calculated from 500 samples drawn from Design 4.

The results are given in Figures 2.10 to 2.12 and Table 2.5. We see that both packages manage to predict \( g(.) \) with again a better fit for \texttt{mgcv}. Clearly, \texttt{gamlss}
barely fits \( h(\cdot) \) with important errors and a very thick confidence band that make it unreliable, whereas \texttt{mgcv} exhibits a much better fit. However, when the sample size increases, the problems encountered by the \texttt{gamlss} package tend to be mitigated.

It should be finally remarked that there exists also the possibility to estimate varying-coefficient models in the BayesX R package. This is a software developed by Belitz, Brezger, Kneib, Lang and Umlauf (2013). It was originally designed to mainly deal with generalized additive models, generalized mixed additive models, generalized geo-additive mixed models, geographically weighted regression, etc. But since a while it includes also special cases of varying-coefficient models. BayesX considers three inferential procedures: Full Bayesian (FB) inference based on MCMC simulation techniques (the function is \texttt{bayesregobjects}), mixed model based estimation corresponding to an Empirical Bayesian (EB) inference or a penalized likelihood (the function is \texttt{remlregobjects}), and inference based on penalized least squares (the function is \texttt{bstepwiseregobjects}). In brief, in the
Figure 2.10: Design 4, 90% simulation intervals for sample size $n = 50$ for $g$ on the left, $h$ on the right.

case of the FB approach, the prior of the coefficient functions are assumed to be random with appropriate hyper-priors whereas in the EB approach they are assumed to be unknown parameters. Let us consider these procedures in the specific case of varying-coefficient models which can be divided in three subgroups: models with continuous variables as effect modifiers, models with spacial coordinates, and models with unordered group indicators (say, random slopes). For the inferences based on the FB or on the EB approach, the three types are implemented in the same way. However, in our study we are more concerned with continuous variables. This case can be estimated in (at least) three different ways: 1st \texttt{rw1} or 2nd \texttt{rw2} order random walk, P-spline with 1st \texttt{psplinerw1} or 2nd \texttt{psplinerw2} order random walk penalty – or with a ‘seasonal prior’. For the case of doing inference based on a penalized least squares approach, two methods are implemented for continuous regressors: one considering P-splines with 1st or 2nd order difference penalty, and the another one with zero degree P-splines. For non-Bayesian
Figure 2.11: Design 4,90% simulation intervals for sample size \( n = 100 \) for \( g \) on the left, \( h \) on the right.

Statisticians and therefore for almost all empirical economists it is very hard to see how to choose between this impressive set of options, and to set the parameters. Results certainly also depend on the prior choice and simulation method chosen. Furthermore, without going into the source codes, one can only obtain figures with Bayesian confidence intervals but no predictions at given points, nor numerical results on these. All this together makes it a bit difficult to compare it with the above mentioned packages and methods. It would be, however, interesting to study this direction deeper as some small simulation studies (not presented here) seem to indicate a quite competitive performance.
2.5 Conclusion

First the problem of endogenous coefficients is presented. For them we show the failure of the OLS and usual IV approaches when estimating an average effect or return. Therefore the use of varying-coefficient models is suggested because of their potential to handle heterogeneous returns. One should attempt to model directly the coefficients as a function. This is partly underlined by some example from empirical economics. We have explicitly shown the inadequacy of traditional econometric inference to estimate an average effect when the coefficients are endogenous. We then propose alternative modeling approaches by using varying-coefficient models. This can be extended to IV estimation to overcome the problem of remaining endogeneity.

Some simulation studies were performed to first show the failure of classical methods, and then in Section 2.4 some larger simulation studies to compare the per-
formance of existing varying-coefficient model estimators. More specifically, we have presented several R packages commonly available to implement the discussed varying-coefficient models.

This article has highlighted the potentiality of varying-coefficient models to estimate individual or average effects of regressors while usual econometric approaches fail because of considering endogenous coefficients. Regarding existing estimation methods, we have provided an overview of the literature on estimators for varying-coefficient models and their available implementations in R. At this point, for sake of generality, we have proposed solutions to deal with endogenous coefficients in several specifications. Nevertheless, we hope that our contribution will motivate further researches on this track.
Bibliography


Chapter 3

On Semi-parametric Models for Evaluating Endogenous Effects Correlated with Instruments through Unobservable Abilities

3.1 Introduction

Identifying and estimating the effects of a variable on some outcome is crucial in econometrics. In non-experimental situations, the variable of interest, called treatment, is the result of an individual choice that leads to a confounding issue when self-selection is not taken into account. A standard approach to overcome this problem is to control for the individual choice by including relating covariates to exogenize the treatment. However, despite the current rich data environment, a lot of information on individual abilities relating to the self-selection remains unobservable. In those cases, the usual identification of causal effects relies on the observation of instrumental variables. A valid instrument should be correlated only
with the treatment when conditioning on the covariates. However, this condition is often not fulfilled in practice as many pertinent instruments are correlated with the unobserved heterogeneity of the outcome through unobservable abilities.

Valuable instruments require a sufficiently high correlation with the treatment variable. Such instruments are likely part of the data generating process as determinants of the treatment. In this context, the Generalized Roy model (Heckman and Vytlacil, 1999) endows the Rubin (1974) potential outcomes model with a structural selection determined by the instruments and covariates. In a non-parametric point of view, a valid instrument should be independent of the potential outcomes when conditioning on the covariates. That is the so-called Conditional Independence Assumption of Instrumental Variables (CIA-IV). Following the causal approach of the structural model, we endow the latter with a coherent Directed Acyclic Graph (DAG) while taking into account unobservable abilities correlated with the instruments. At the light of this graph, we show that if unobservable abilities interact with or depend on the instrument (once having conditioned on covariates), then the CIA-IV is violated for both the selection rule and the potential outcomes.

We propose to consider a mean independent additive unobservable term representation of the selection equation that is assumed to capture all spurious associations between the instruments and the unobservable heterogeneity of the outcome. In other words, the violation of the CIA-IV takes the form of a Conditional Mean Restriction (CMR) on the potential outcomes once conditioned on covariates and the unobservable yet predictable additive error term of the considered selection rule representation. In this context, both the Rubin (1974) model and the proposed generalization to the continuous treatment are expressed as a varying-coefficient model (Hastie and Tibshirani, 1993). From the relating non-parametric slope function we can identify causal effect and policy-relevant parameters. the latter is expressed in terms of the instrument when it is under the control of the policymaker.
The illustrative application we propose is based on the data provided by Maynard and Qui (2009) and Gruber and Yelowitz (1999). In their articles, they have assessed the deterrent impact of the public health insurance program, Medicaid, on American household savings. We consider a continuous version of our model that takes into account the self-selection of the household through abilities represented by the unobserved health status and its apparent correlation with the used instrument. The obtained results suggest that the deterrent effect has been underestimated in the literature.

In the binary treatment effect context, we can reconsider the latent variable selection rule by Heckman and Vytlacil (1999) that express the error terms as a quantile of unobservable abilities. In this case, we derive and identify an average effect parameter on the indifferent-to-participation sub-population. From the arc counterpart of this parameter, we can evaluate the impact of participating due to a change in the unobservable abilities quantile induced by a policy implementation that shifts the instruments and the covariates.

The structure of this article is the following. Section 2 introduces the Generalized Roy Model supplemented with unobservable abilities that yield the violation of the CIA-IV. Section 3 motivates the validity of CMR that yields to a semi-parametric representation of the Generalized Roy Model and its generalization to the continuous treatment from which relevant parameters can be identified. Section 4 proposes point-wise estimation strategies and their related confidence intervals based on a bootstrap procedure. Section 5 contains an empirical application based on the Medicaid data. Section 6 concludes. Finally, Section 7 contains similar considerations related to the binary treatment case.
3.2 Unobservable Abilities and Their Influence on The Instruments

This section investigates the role that unobservable abilities play in the identification of the effects in presence of endogeneity. In particular, we consider a Generalized Roy Model (Heckman 2010) that bridges the gap between the structural and counterfactual approach in econometrics. This model considers an endogenous binary treatment with a selection on the "unobservables" through abilities that invalidate the CIA-IV required for the usual identification of effect parameters.

Let be the following Generalized Roy Model with a micro-economically interpretable selection rule similar to Heckman (2010) where the instruments are introduced as cost shifters:

\[ Y_i = \alpha_i \cdot Y^0_{i} (X_i, W_i, U^0_i) + \beta_i \cdot Y^1_{i} (X_i, W_i, U^1_i) - Y^0_{i} (X_i, W_i, U^0_i) \cdot D_i, \]

\[ D_i = 1 \{ E[\beta_i - C_i(Z_i) | I_i] > 0 \}, \]

\[ I_i = (A_i, X_i, Z_i) \quad i = 1, ..., n, \quad \text{(3.1)} \]

with \( \alpha_i = Y^0(X_i, W_i, U^0_i) \) and \( Y^1(X_i, W_i, U^1_i) \) being the outcome when individual \( i \) is respectively untreated and treated. \( i \) decides to participate when her expected return of participation \( \beta_i \), reduced by the cost of change \( C_i(Z_i) \), is positive given her information set when the decision is taken, namely \( I_i \). The latter contains the control variables \( X_i \) taking their values in \( \mathbb{R}^K \), the instruments \( Z_i \), and the vector of abilities \( A_i \) that are observed only by individual \( i \). We generalize the covariates by including the variables \( W_i \) taking their values in \( \mathbb{R}^L \) that are unobservable by or irrelevant to Individual \( i \) before her participation decision yet impacting on her outcome. These latter variables are covariates like \( X \) but affect only on the outcome as the latter is often observed later after the policy implementation. Finally, \( U^0 \) and \( U^1 \) are unobserved heterogeneity of the outcomes summarized in a single random variable for each potential outcome.
Unobservable abilities may interact with instruments and the usual covariates like gender, age, socio-economic background and culture proxy variables. As a result, we consider a general expected utility latent variable as a non-parametric function as follows:

$$E[\beta_i - C_i(Z_i)|I_i] = \mu(A_i, X_i, Z_i).$$  \hspace{1cm} (3.2)

Nowadays, causal graphs are very popular in statistics to represent the relationship between the variables of a non-parametric model. In particular, The properties of directed acyclic graphs (DAG), their utility and use have been abundantly documented by Pearl (1999). Even if the outcome equation of model The Generalized Roy Model (3.1) is semi-parametric, we can represent the relevant relationship of the non-parametric potential outcomes, and the generalized latent variable selection rule (3.2) by the DAG depicted in Figure 3.1. Solid edges represent the effect of an observable variable like $X \rightarrow Y^d$. Uni-directed dashed edges describe the effect of an unobservable variable like $A \rightarrow U$. Bi-directed dashed edges denotes broader associations like correlations most often coming from a common factor structure. In this case, $A \leftrightarrow Z$ denotes for instance $A \leftarrow O \rightarrow Z$ where $O$ is unobservable.

Path $Z \rightarrow D \rightarrow Y^d$ is policy-relevant when $Z$ is at least partially under the control of the policymaker. In this case, all other paths from $Z$ to $Y^d$ are spurious. The usual way to deal with these confounding paths in a non-experimental data environment is to block them by conditioning on control variables, namely $(W, X)$ in our case. However, the indirect association between $Z$ an $Y^d$ through $A$ and $U$ cannot be blocked by conditioning $(W, X)$ as depicted by the DAG in Figure 3.1. Furthermore, since $A$ is unobservable, we cannot block this spurious path by conditioning on it. A solution is to control for $A$ through a variable we can identify and thus predict. In this article, we implement this solution.

A correlation between unobservable abilities and relevant instruments is realistic
Figure 3.1: DAG representation of the Generalized Roy Potential Outcome ($Y^d$) Model with unobserved abilities ($A$) correlated with the instruments $Z$ in many situations. Indeed, an instrument which is a strong determinant of the selection is likely correlated with unobservable abilities. As an example, Card (1995) considers the proximity to college as an instrument for the years of schooling when investigating the impact of the latter on the student future earnings. Carneiro and Heckman (2010) have shown that the abilities of the students living close to a college are higher than in other places given the covariates. These skills will probably help them to earn a higher income in the future.

### 3.3 Semi-parametric Representations and Identification

In this section, we consider a way to deal with the violation of the CIA-IV by assuming that all spurious associations can be captured the additive mean independent error term obtained from the selection rule. In this case, Model (3.1) can
be represented by a semi-parametric specification and generalized to the continuous
treatment by modeling the parameters of interest with non-parametric functions.

We consider a semi-parametric representation that enables us to net out the effect
of all spurious associations between Z and Y that are captured by an unobservable
yet predictable additive term. On this purpose, we consider the following fully non-
parametric selection rule:

\[
D_i = \mathbb{1}\{E[\beta_i - C_i(Z_i)|I_i] > 0\} = \mathbb{1}\{\mu(A_i, X_i, Z_i) > 0\},
\]

\[= D(A_i, X_i, Z_i), \quad i = 1, ..., n. \]

Then, as pointed out by Matzkin (2003) and implemented by Moffitt (2008), we
can represent \(D(A_i, X_i, Z_i)\) as follows:

\[
D_i = E[D_i|X_i, Z_i] - \{E[D_i|X_i, Z_i] - D(A_i, X_i, Z_i)\},
\]

\[= P(X_i, Z_i) - V(A_i, X_i, Z_i) \quad \text{3.3} \]

such that \(E[V_i|X_i, Z_i] = 0\) for all \(i\) by construction. In other words, we represent
the selection rule by a non-parametric propensity score plus an additive error term \(V\)
that is predictable. \(V(A, X, Z)\) can be interpreted as a deviation of the probability
to participate from the observable mean behaviors, represented by the propensity
score \(P(X, Z)\), due to an individual private information on abilities \(A\). Going back
to the Card (1995) example; if a student has higher skills to succeed at school,
then, she would be more likely to enter the college than other children with similar
observable characteristics.

Now, we introduce a crucial condition on \(V\) that is required for the identification
of the parameters presented at the end of the chapter. All confounding association
between \(Z\) and \(Y^d\) that are not controlled by conditioning on \(X\) and \(W\) can be
captured by \(V\). In other words, the violation of the CIA-IV takes the form of a
Conditional Mean Restriction (CMR) on the potential outcomes as follows:

\[ E[Y^d|V(A, X, Z), X, W, Z] = E[Y^d|V(A, X, Z), X, W], \quad d = 0, 1. \] 

In other words, the unobservable abilities impact on the future outcomes only through a deviation of the probability to participate from the propensity score. In this case, \( V \) has a crucial property; unlike \( U_d \), it is identifiable, and its prediction can be used as a control variable.

Intuitively, we require that the association between \( A \) and \( Y^d \) is fully channeled through \( V \) when any variation in \( A \) is unambiguously represented by a variation in \( V(A, X, Z) \). In mathematical terms, that underlies the invertibility of \( V(., X, Z) \). Formally,

**Proposition 3.** For \( i = 1, ..., n \) such that \( V_i = V(., X_i, Z_i) \) satisfying Model (3.1) endowed with the DAG depicted on Figure 3.1 with selection rule provided by Equation (3.3).

If \( V(., X_i, Z_i) \) is invertible, then \( E[Y^d|V_i, W_i, X_i, Z_i] = E[Y^d|V, W, X] \).

A proof is given in the appendix.

Now, since the outcome equation of Model (3.1) can be expressed as a random coefficient model, we extend the CMR on potential outcomes (3.4) to \( \alpha \) and \( \beta \) as follows:

\[ E[\alpha_i|P(X_i, Z_i) - D_i, W_i, X_i, Z_i] = E[\alpha_i|P(X_i, Z_i) - D_i, W_i, X_i], \]
\[ E[\beta_i|P(X_i, Z_i) - D_i, W_i, X_i, Z_i] = E[\beta_i|P(X_i, Z_i) - D_i, W_i, X_i]. \] 

(3.5)

These CMR based on the coefficient representation of the Rubin model will enable us to extend them to coefficients of the continuous treatment model. The
parameters proper to the binary case are of interest in treatment effect and policy evaluation literature. However, we do not consider the related parameters and their identification here as we are interested only in providing a counterfactual justification of the continuous case based on the potential outcome specification. As a result, the binary case considerations are provided in the appendix.

Similarly to a binary treatment, a continuous choice variable like that quantity of a good or service depends most likely to unobservable abilities. In this case, new average effect parameters have to be propose. The continuous treatment we consider can be interpreted as an extension of the non-parametric model proposed by Imbens and Newey (2009) when the violation of the CIA-IV occurs through the self-selection on unobservable abilities.

Even if the generalization of the Rubin model to continuous treatment may appear intuitive and natural, we present here a formal argument. We introduce the usual notation $S$ for the non-binary treatment variable. For the sake of clarity, we first consider the Rubin model with three potential outcomes. Thus, similarly to Angrist and Imbens (1994), we obtain

$$Y_i = Y_i(0) + [Y_i(1) - Y_i(0)]I(S_i \geq 1) + [Y_i(2) - Y_i(1)]I(S_i \geq 2),$$

$$S_i \in \{0, 1, 2\}, \quad i = 1, \ldots, n.$$

The generalization to the model with $S$ discrete and $J$ potential outcomes is

$$Y_i = Y_i(0) + \lim_{J \to \infty} \sum_{j=1}^{J} [Y_i(j) - Y_i(j-1)]I(S_i \geq j),$$

$$S_i \in \mathbb{N}, \quad i = 1, \ldots, n.$$

The next proposition enables us to express this potential outcome in the terms of varying-coefficient model.

**Proposition 4.** For $i = 1, \ldots, n$ such that $Y_i(j)$ is the value that the potential outcome of individual $i$ take at the level of intensity $j$ of the treatment and $S_i$ the
value of the treatment that $i$ chooses, if $Y_i = Y_i(0) + \lim_{J \to \infty} \sum_{j=1}^{J} [Y_i(j) - Y_i(j - 1)] \mathbb{I}(S_i \geq j)$, then

$$Y_i = Y_i(0) + B_i(S_i),$$

where $B_i(.)$ is a random function such that $B_i(0) = 0$ a.s..

A proof is given in the appendix.

To establish a conformity with the perspective of the Generalized Roy Model (3.1), we simply assume that $B_i(S_i) = B := \beta_i S_i \alpha_i := Y_i(0)$ with the following representation:

$$Y_i = \alpha_i(W_i, X_i, U_i^0) + \beta_i(W_i, X_i, U_i) S_i,$$

$$S_i \in \mathbb{R}_+, \quad i = 1, \ldots, n,$$

where $U_i$ is a univariate random variables that summarizes the unobservable heterogeneity of the effect for individual $i$. All other variables are defined similarly to Model (3.1).

$Y_i$ is maximized with respect to $S_i$ by individual $i$ given her information set $I_i = (A_i, X_i, Z_i)$. In other words, we have the following structural model:

$$Y_i = \alpha_i + \beta_i S_i,$$

$$S_i = \arg\max_{s \in \mathbb{R}_+} E[\beta_i s - C_i(s, Z_i)|A_i, X_i, Z_i],$$

$$= g(X_i, Z_i) - V(A_i, X_i, Z_i), \quad E[V_i|X_i, Z_i] = 0.$$

(3.6)

Similarly to the CMR assumed in the binary case given by equation (3.5), we assume the following,

$$E[\alpha_i|g(X_i, Z_i) - S_i, W_i, X_i, Z_i] = E[\alpha_i|g(X_i, Z_i) - S_i, W_i, X_i],$$

$$E[\beta_i|g(X_i, Z_i) - S_i, W_i, X_i, Z_i] = E[\beta_i|g(X_i, Z_i) - S_i, W_i, X_i].$$

(3.7)
Thus, we can apply the CMR (3.5) on the outcome equation of Model (3.6) to obtain

\[
E[Y_i|S_i, X_i, Z_i, W_i] = E[\alpha_i|S_i, W_i, X_i, Z_i] + E[\beta_i|S_i, W_i, X_i, Z_i]S_i ,
\]

\[
= E[\alpha_i|S_i = g(X_i, Z_i) - V_i, W_i, X_i, Z_i] + E[\beta_i|S_i = g(X_i, Z_i) - V_i, W_i, X_i, Z_i]S_i ,
\]

\[
= E[\alpha_i|V_i = g(X_i, Z_i) - S_i, W_i, X_i, Z_i] + E[\beta_i|V_i = g(X_i, Z_i) - S_i, W_i, X_i, Z_i]S_i ,
\]

\[
= E[\alpha_i|g(X_i, Z_i) - S_i, W_i, X_i] + E[\beta_i|g(X_i, Z_i) - S_i, W_i, X_i]S_i ,
\]

(3.8)

where the last equation comes from CMR assumed on the potential outcomes (3.7).

Finally, we obtain the following semi-parametric equation system:

\[
E[Y_i|V_i = g(x, z) - s, W_i = w, X_i = x] = \alpha(g(x, z) - s, w, x)
\]

\[
+ \beta(g(x, z) - s, w, x)s
\]

\[
E[D_i|X_i = x, Z_i = z] = g(x) ,
\]

(3.9)

where the outcome equation is a varying-coefficient model (Hastie and Tibshirani, 1993) such that \(\alpha(g(x, z) - s, w, x)\) and \(\beta(g(x, z) - s, w, x)\) are the non-parametric varying-coefficient functions.

In this context, we propose several parameters. If the policy relevance is of interest as \(Z\) is chosen by the policy maker, then we can consider an average effect of a particular level of treatment \(S_i = s\):

\[DMTE(s, w, x, z) := E[\beta_i|V_i = g(x, z) - s, W_i = w, X_i = x] = \beta(g(x, z) - s, w, x) .\]

In particular, a more synthetic average effect can be obtained by evaluating the previous parameter at the sample average treatment, \(s = \bar{s}\). Alternatively, if the causal effect is of interest, \(V_i = v\) can be integrated out the obtain the following average effect:

\[ATE(w, x) := \int_{\mathbb{R}} E[\beta_i|V_i = v, W_i = w, X_i = x]f(v)dv = \int_{\mathbb{R}} \beta(v, w, x)f(v)dv .\]
The identification issue is related to the representation of stochastic and estimable quantities; $E[Y|V = g(x, z) - s, W = w, X = x]$ and $E[S|X = x, Z = z]$ of an unique unobservable and theoretical quantity; respectively $\pi(g(x, z) - s, w, x)$ and $g(x, z)$. In this case, an usual relevance assumption of instruments is sufficient to make $E[S|X, Z]$, a non-trivial random variables required for the identification of $g(x, z)$ as well as, $g(x, \cdot)$ a non-constant function required for the identification of $\pi(g(x, z) - s, w, x)$. Indeed, this condition is required to introduce a variation in the first argument $g(x, z) - s$ that is free from the variation in the third argument $x$ when $s$ is fixed. Once $\pi(g(x, z) - s, w, x)$ is identified, a possible confounding between $\alpha(g(x, z) - s, w, x)$ and $\beta(g(x, z) - s, w, x)s$ may yield to an second identification issue. The latter is solved thanks to the varying-coefficient structure that enables to net out the identification of $\alpha(g(x, z) - s, w, x)$ from the identification of $\beta(g(x, z) - s, w, x)s$ when $s = 0$. This identification would have been impeded in a more general additive model.

### 3.4 Estimation and Confidence Intervals

This section is devoted to the estimation of the DMTE parameter defined previously. The proposed strategy is a multi-stage estimation where every single step can be easily implemented with most of the statistical and econometric software by plugging in the estimator of the previous step. Different methods are proposed depending on the dimension of $X$ and $W$. Finally, we propose a bootstrap strategy to build a confidence interval for the obtained point estimates.

The first stage concerns the estimation of $g(x, z)$. If the dimension of $x$ and $z$ is small, then we can implement a local polynomial regression with a kernel function. However, in most cases, the dimension of $X$ is too high to sustain a full non-parametric estimation. An intermediate alternative to mitigate the curse of
dimensionality would be to assume additive splines and run a standard OLS estimation.

Kernel regressions are nowadays very popular in non-parametric econometrics because they produce natural empirical counterparts of the conditional expectations evaluated at a specific value. However, the backfitting algorithms for kernel estimation are seldom implemented by the statistical software. Furthermore, since kernel methods are local estimations, they are not well designed for a large interpolation of $\beta$ from $\hat{g}(x, z)$ to $\hat{g}(x, z) - s$. As a result, we rather consider spline methods that are widely used in statistics but seldom implemented in applied econometrics.

Spline methods are local like kernel ones and, thus, perform similarly. The latter can be seen as a polynomial of order 2 added with local third order terms defined locally between knots. The global component of splines enables more reliable large interpolations the kernel methods. As the sample size increases, the distance between the knots should decrease, similarly to $h$ in kernel estimation, to improve the inference. That leads to smoothing splines where the knots are located at each unique observed value of the regressors ruling out any arbitrary selection of knots. The relating over-parametrization is balanced by minimizing an objective function (least squares or log-likelihood) with a penalty on the global curvature.

It has been shown that a reduced number of knots, usually at each 10, 15 or 20 observations is sufficient to make equivalent any arbitrary selection of their location. Furthermore, a much lower number of knots reduces an unnecessary complexity and thus should improve the estimation. From this point of view, B-splines are well designed in the univariate case and can be generalized by taking the tensor product of the marginal (univariate) B-splines. In this case, the non-
Parametric functions can be expressed as follows:

\[
\alpha(v, x, z) = A(v, x, z)^T a = \{A(v) \otimes_l A_l(w_l) \otimes_k A_k(x_k)\}^T a ,
\]

\[
\beta(v, x, z) = B(v, x, z)^T b = \{B(v) \otimes_l B_l(w_l) \otimes_k B_k(x_k)\}^T b ,
\]

where \(A(v, x, z)\) and \(B(v, x, z)\) are the raw matrices of B-splines.

Since the number of parameters remains relatively high, a penalization for the curvature is considered to prevent over-fitting (Eilers and Marx, 1996). As a result, this method is often called “P-splines.” Wood (2011) proposes a computationally fast algorithm implemented by the R package, mgcv. Under some distribution family assumption (Gaussian by default) characterized by the B-Splines, the estimates of \(a\) and \(b\) are obtained by penalized maximum likelihood method as follows:

\[
(\hat{a}, \hat{b}, \lambda^*) = \arg \max_{a, b, \lambda} l(Y|\hat{V}, X, Z, D; a, b)) - \lambda \int \int \int (a^T H[A(v, x, z)]a + b^T H[B(v, x, z)]b)dvdwdx ,
\]

where \(l\) is the log-likelihood function, \(H[A(v, x, z)]\) and \(H[B(v, x, z)]\) are respectively the Hessian matrix of \(A(v, x, z)\) and \(B(v, x, z)\), and \(\lambda\) is the smoothing parameter.

Wood (2011) considers a Penalized Iteratively Reweighted Least Squares (PIRLS) algorithm to solve this optimization problem. At each iteration, the optimal \(\lambda\) is obtained by (generalized) cross-validation with \(a\) and \(b\) that computed in the previous iteration. The algorithm should converge to the optimum \((\hat{a}, \hat{b}, \lambda^*)\). In the specific case of varying-coefficient models, the mgcv package exhibits good finite samples properties compared to other known R algorithms (Sperlich and Theler, 2015). As a result, this package will be used in the application discussed in the next section.
Once $\tilde{a}$ and $\tilde{b}$ are obtained, we predict $MPRTE$ at the specific values $(\tilde{g}(x, z), w, x)$. The reliability of this prediction is measured by a confidence interval. The latter can be computed using a bootstrap method that has proven to be more precise than the usually approximated interval confidences based on the asymptotic normality. The advanced and recent bootstrap methods involve Monte-Carlo simulations to generate error terms and obtain replicated estimates of the parameters of interest. In our case, the considered multi-stage estimation introduces some heteroscedasticity at the outcome equation that can be handled by the wild bootstrap (Mammen, 1993). Furthermore, the use of an instrumental variable in the proposed estimation methods requires replicating the estimates of both the selection equation and the treatment. In this case, we consider the following bootstrap strategy:

1. For $r = 1, ..., rep$ and $i = 1, ..., n$.
2. Generate $U_r^i \sim N(0,1)$.
3. Predict $\hat{g}(X_i, Z_i)$ after regressing $S$ on $(X, Z)$.
4. Generate $\hat{V}_i = \hat{g}(X_i, Z_i) - S_i$.
5. Generate $\hat{V}_r^i = \hat{V}_i U_r^i$.
6. Generate $S_r^i = \hat{g}(X_i, Z_i) - \hat{V}_r^i$.
7. Predict $\hat{g}'(X_i, Z_i)$ after regressing $S'$ on $(X, Z)$.
8. Predict $\hat{\alpha}(\hat{V}_i, X_i, W_i)$ and $\hat{\beta}(\hat{V}_i, X_i, W_i)$ after regressing $Y$ on $(\hat{V}_i, X_i, W_i)$.
9. Generate $\hat{\sigma}_i = Y_i - \hat{\alpha}(\hat{V}_i, X_i, W_i) - \hat{\beta}(\hat{V}_i, X_i, W_i) S_i$.
10. Generate $\varepsilon_r^i = \hat{\sigma}_i U_r^i$.
11. Generate $Y_r^i = \hat{\alpha}(\hat{g}'(X_i, Z_i) - S_r^i, X_i, W_i) + \hat{\beta}(\hat{g}'(X_i, Z_i) - S_r^i, X_i, W_i) S_r^i + \varepsilon_r^i$. 

81
12. Predict $\hat{\beta}(\hat{g}(x, z) - s, x, w)$ after regressing $Y^r$ on $(\hat{g}^r(X, Z) - S^r, S^r, X, W)$.

Finally, we can build a 90% confidence interval of the estimated MPRTE from the outer envelope obtained by excluding 5% both highest and lowest values of $\hat{\beta}(\hat{g}^r(x, z), x, w)$.

3.5 The Empirical Illustration

In this section, we apply our models to the data provided by Gruber and Yelowitz (1999) and supplemented by Maynard and Qui (2009) who investigate the deterrent impact of the public health insurance program, Medicaid, on the savings of the American households in a random sample of 40442 observations. Similarly, we consider their Medicaid generosity measure that summarizes both the eligibility to the insurance and its expected subsidies. That is coherent with the continuous treatment model, presented in the previous section, where the household health status plays the role of the unobservable abilities.

The Medicaid is a first-dollar (100% coverage) public health insurance implemented at the level of each state. It has been extended by the Patient Protection and Affordable Care Act, so-called ”Obamacare.” Similar to Gruber and Yelowitz (1999), Maynard and Qui (2009) consider the Medicaid Eligible Dollars (MED) as a measure of the Medicaid generosity. The MED is defined as the present value of the future expected subsidies of the household members computed with the eligibility probabilities being conditioned on the family members characteristics, their level of income and assets. The considered outcome variable is the net worth $\ln(Y)$ which is the natural logarithm of the valued sum of financial assets, home, vehicles and business equities net of debt. This semi-log specification enables us to consider the effect of the generosity regarding a relative variation of the savings.
The self-selection endogeneity of the MED occurs in two ways. First, the households together affect the legislation of their state on the level of MED through lobbies. Second, the households can individually decide to benefit from the asset test Medicaid subsidies by lowering their level of savings. To exogenize the MED from the lobbying effect, Gruber and Yelowitz (1999) and then Maynard and Qui (2009) include a large set of covariates. The latter contains binary variables, namely state, year and their interactions as well as categorical variables relating to the size of families, age–gender and age–education groups. Regarding the endogeneity coming from the lowering of the savings level to obtain the asset tests eligibility, an IV approach is considered. Similarly to Gruber and Yelowitz (1999), Maynard and Qui (2009) built an exogenous measure of coverage, the Simulated Medicaid Eligible Dollars (SIMMED). The latter is similar to the MED except that the eligibility probabilities are conditioned only on non-manipulable household characteristics: education, members age, state and year. These variables are considered as exogenous by Gruber and Yelowitz (1999) who argue that they determined externally to the saving decision and, then, are not subject to self-selection of households.

Nevertheless, the SIMMED depends also on the unobserved health status of the households from the demand side of health care. However, since the data do not contain any direct observations of the health of the households, the role of the latter has been somehow underestimated. The behavior of the households facing an eligibility decision is undoubtedly affected by their health status. As a result, a theory that considers the role of health as an unobservable ability is required. Furthermore, the policymaker has also an influence on the SIMMED from the supply side of health care. As a result, this context also justifies the interest to estimate policy-relevant parameters.

Maynard and Qui (2009) find that the households are willing to decrease their level of assets by 2.5% on average to obtain 1000 more eligible dollars (MED). However,
this result has been obtained by the two-stage least squares (2SLS) method that produces biased estimates of the average effect when the treatment is correlated with effect as it is well established in policy evaluation literature. Maynard and Qui (2009) also implement a quantile regression by modeling the heterogeneous impact in terms of net worth quantiles. This approach is informative from an economic and policy perspective. However, it is not crucial from an identification point of view as long it is not designed to deal with a violation of the CIA-IV.

Now, let us consider the adaptation of Model (3.9) to the Medicaid context. Since we have 342 covariates, an additive structure similar to Specification (3.15) is required. Furthermore, apart from the age of the household head (Age), all covariates are either binary or ordinal variables with 3 categories. As a result, we specify a full linear structure on the covariates while considering the age as a continuous variable that is included in the non-parametric functions as follows:

\[
E[\ln(Y_i)|g(X_i, Age_i, SIMMED_i) - SIM_i = g(x, a, z) - s, Age_i = a, X_i = x] = \alpha_0 + x^T\alpha_x + ax^T\alpha_{ax} + \alpha_v(a, g(x, z) - s) + [\beta_0 + x^T\beta_x + ax^T\beta_{ax} + \beta_v(a, g(x, z) - s)]s . \tag{3.10}
\]

Since the age does not play any particular role compared to other covariates, we consider the following change of notation:

\[
Y^0(a, s, x, z) = Y^0(s, x, z) ,
\]
\[
DMTE(a, s, x, z) = DMTE(s, x, z) .
\]

Figure 3.2 summarizes the main results obtained by the wild bootstrap proposed in the previous section. The non-parametric curves estimating \( Y^0(\bar{s}, \bar{x}, z) \) and \( Y^0(\bar{s}, \bar{x}, z) \) are obtained by their prediction at equidistant points on \( z \) while evaluating at the mean of \( s \) and \( x \). Then, the confidence intervals are obtained at the 0.05 and 0.95 quantiles of 1000 replicates at each point-wise prediction.
Figure 3.2: Predicted Non-parametric both intercept $Y^0$ and slope DMTE as functions of $z$ in Model (3.12).

Before analysing the main results, we consider the histogram of the instrument $Z:=\text{SIMMED}$ depicted on Figure 3.3. The latter suggests that most households should expect about $z = 5000$ SIMMED in case of exogenous eligibilities (independently on their decision to lower their savings to obtain it). From a statistical point of view, we observe very little households larger than $z = 15000$ or $z = 20000$ SIMMED. As a result, we should expect unreliable results for those values due to a relating lack of information. This is confirmed by Figure 3.2 that depicts very large confidence interval for high values of $Z$.

On the left part of Figure 3.2, we observe an increase of the non-eligible Medicaid household savings at least up to $z = 15000$ SIMMED suggesting that non-insured households may build some precautionary savings to face potential health expenditure. On the left part, we observe a clear decreasing trend suggesting that Medicaid insured households are willing to decrease their savings as they expect higher ex-
Figure 3.3: Histogram of the instrument $Z:=$SIMMED.

Penditure costs given their health status. This decrease is at a minimum of 7.5% to about 25% for $z = 25000$ SIMMED most likely associated with a low level of health.

Here it is crucial to note that the 2SLS estimate is significantly higher than the DMTE whatever the level of $z$ SIMMED. That implies that the average deterrent effect of the Medicaid on savings has been underestimated by the usual 2SLS. We have drawn a similar conclusion by considering other values of $x$ and $s$.

Regarding policy-relevance recommendations based on our results, Figure 3.2 can provide some insights. The DMTE seems the highest at values of lower than $z = 5000$ SIMMED. Therefore, a public insurance that would cover only healthier households, when associated with a low level of SIMMED, would probably minimize the deterrent effect on savings even if the more vulnerable households would remain uninsured.
3.6 Conclusion

In this essay, we have justified and discussed the identification and the estimation of endogenous effect parameters when the usual conditions assumed on the instruments are not fulfilled due to unobservable abilities. That is justified by a Generalized Roy Model framework that includes the role played by unobservable abilities.

In this context, the violation of the usually assumed conditional independence of the instruments is due to a confounding correlation between the instruments and the unobserved heterogeneity of the potential outcomes through unobservable abilities. The self-selection rule provides a structural approach in terms of an individual maximization problem that defines the instruments as a determinant of the endogenous treatment. In many situations, the individuals base their decision on abilities that are unobserved in many data sets. In those cases, the unobservable abilities are determinants of the unobserved heterogeneity of both the outcome and the treatment while being correlated with the instruments. That impedes the usual identification of pertinent parameters.

As a remedy, we have proposed a semi-parametric representation of the Generalized Roy Model and its continuous treatment generalization that models the violation of the conditional independence assumption as a conditional mean restriction on the potential outcomes. In this context, we provide a policy-relevant parameter depending on the instrument that is under the control of the policy maker as well as a causal parameter that can be obtained by marginal integration.

We have provided a natural continuous treatment generalization. In this context, the proposed parameters depend on the intensity of both the treatment and the unobserved abilities through the instruments. We have illustrated the theory by running the continuous treatment model with the Medicaid data where the
household health status plays the role of unobserved abilities. We have found the two-stage least squares estimation provided in the literature has under-estimated the deterrent effect of the Medicaid on the household savings. Besides, we found that the deterrent effect increases with the exogenous subsidies probably associated with a decrease in the health status.

Appendices

A. The Binary Treatment Effect Parameters and Their identification

This part of the appendix is devoted to the identification and estimation of the binary treatment effect parameters that are of interest in impact and policy evaluation literature. On this purpose, we introduce the parameters we will derive from the CMR semi-parametric representation of Model (3.1). Then, we consider a constructive and point-wise identification of these parameters based on the non-parametric varying-coefficient structure of the outcome equation. This identification strategy is based assumptions that are common in the treatment effects literature.

The parameters we propose do not manage to identify the nowadays popular Local Average Treatment Effect of the compliers. However, they have an interesting policy-relevant feature as they enable to express the average effect with respect to the instruments, through the propensity score, when some of them are under the influence of the policymaker. Alternatively, we can also identify the impact of participation due to a change in unobservable ability quantiles. If a causal effect parameter is considered as more relevant, then the usual Average Treatment Effect can be obtained by integrating out the propensity score.
The following model is the binary treatment counterpart of Model (3.6) based on the CMR assumed on coefficient given by equations 3.5.

\[ Y_i = \alpha_i + \beta_i D_i, \]
\[ D_i = P(X_i, Z_i) - V_i, \quad P(X_i, Z_i) = E(D_i|X_i, Z_i) \]

CMR: \[ E[\alpha_i | P(X_i, Z_i) - D_i, X_i, W_i] = E[\alpha_i | P(X_i, Z_i) - D_i, X_i, W_i], \]
\[ E[\beta_i | P(X_i, Z_i) - D_i, X_i, W_i, Z_i] = E[\beta_i | P(X_i, Z_i) - D_i, X_i, W_i]. \]

(3.11)

Similarly to Model (3.9), we consider the following semi-parametric representation of (3.11).

\[ E[Y | V = P(x, z) - d, W = w, X = x] = \pi(P(x, z) - d, w, x), \]
\[ = \alpha(P(x, z) - d, w, x) + \beta(P(x, z) - d, w, x) \]
\[ E[D | X = x, Z = z] = P(x, z). \]

(3.12)

A standard approach to assess the endogenous binary treatment effects is to consider the Marginal Treatment Effect (MTE) parameter by Heckman and Vytlacil (2005). The usual identification of the latter is based on a latent variable model with an unobservable additive term. Then, the CIA-IV on unobservable heterogeneity is required yet over-assumed to identify the MTE. That can be shown by considering the following semi-parametric representation of the generalized latent variable selection rule (3.2):

\[ \mathbb{I}\{E[\beta_i - C_i(Z_i) > 0 | Z_i]\} = \mathbb{I}\{\mu(A_i, X_i, Z_i) > 0\} \]
\[ = \mathbb{I}\{P(X_i, Z_i) - \left[ P(X_i, Z_i) - \mu(A_i, X_i, Z_i) \right] > 0\} \]
\[ = \mathbb{I}\{P(X_i, Z_i) - \underbrace{U_D(A_i, X_i, Z_i)}_{U_{D_i}} > 0\}, \quad i = 1, ..., n. \]

(3.13)
Heckman and Vytlacil (2005) define an important treatment effects parameter; the so-called Marginal Treatment Effect (MTE). In our context, it is

\[ MTE(u, w, x) = E[\beta_i | U_{Di} = u, W_i = w, X_i = x], \]

The derivation of the latter is tidily based on the selection rule with the following additive unobservable latent term (3.13) when assuming \( U_{Di} \sim U(0, 1) \). The latter assumption enables to summarize the "unobservables" in term of quantiles.

Under the CIA-IV, the Local Instrumental Variable (LIV) parameter (Heckman and Vytlacil, 1999), namely:

\[ LIV(p, w, x) = \frac{\partial E[Y_i | P(X_i, Z_i) = p, W_i = w, X_i = x]}{\partial p}, \]

identifies the MTE as follows:

\[ LIV(p, w, x)|_{p=u} = MTE(u, w, x), \]

where \( u \) is can be evaluated only on the open interval \((0, 1)\) due to the common support assumption of the propensity score required for this identification. From the MTE, all treatments effect parameters can be identified or at least approximated up to the bound of the propensity score. In particular, the LATE (Imbens and Angrist, 1994) is

\[ LATE(w, x, z_1, z_2) = E[\beta_i | P(X_i, z_2) \geq U_{Di} > P(X_i, z_1), X_i = x, W_i = w], \]

This parameter can be seen as an arc version of the LIV parameter. It enables to assess the effects resulting from the policymaker decision shift the instrument from \( z_1 \) to \( z_2 \) on the compliers. The LATE can be derived from the MTE as follows:

\[ LATE(w, x, z_1, z_2) = \frac{1}{P(x, z_2) - P(x, z_1)} \int_{P(x, z_1)}^{P(x, z_2)} MTE(u, w, x) du, \]
Besides, the overall surplus provided by the integral on the MTE can be seen as a specific policy-relevant parameter such as proposed by Heckman and Vytlacil (2001).

When the CIA-IV is not fulfilled, we cannot identify the MTE with the LIV parameter. Indeed, as explained previously, a confounding between the instrument and the ”unobservables” of the selection impedes the identification of an average effect at a specific value $u$ of $U_D$. As a result, we focus on the identification of relevant parameters that can be identified under the CMR. On this purpose, we re-consider the use of the LIV that can be seen as the average effect of the individuals that are indifferent to the participation as follows,

$$LIV(p, w, x) = E[\beta_i | U_{Di} = p, W_i = w, X_i = x],$$

where $p$ is a specific value that the propensity can take. If the indifferent-to-participation sub-population is of interest, then we can define the ”Dual” MTE as follows:

$$DMTE(u, w, x) := E[\beta_i | P(X_i, Z_i) = u, W_i = w, X_i = x].$$

Unlike the LIV, this new parameter can be identified under the violation of the CIA-IV modeled by the CMR of Model (3.11). Similarly to the MTE we can define and identify the Dual Local Average Treatment Effect (DLATE) as the arc counterpart of the DMTE as follows:

$$DLATE(u_1, u_2, w, x) := [\beta_i | u_1 \geq P(X_i, Z_i) > u_2, W_i = w, X_i = x] = \frac{1}{u_1 - u_2} \int_{u_2}^{u_1} DMTE(u, w, x)du,$$

This parameter computes the average effect of individuals who move from non-participation to participation when their level of unobservable abilities quantile shifts from $u_1$ to $u_2$. This parameter is of policy-relevance interest when the policymaker has some influence on the unobservable ability quantiles $U_D(A, X, Z)$ by affecting $X$ and $Z$. 

91
From a causal effect perspective, \( u \) can be integrated out to identify the classical ATE as follows:

\[
ATE(w, x) := E[\beta_i | W_i = w, X_i = x] = \int_0^1 DMTE(u, w, x) f(u) du .
\]

When the instrument is under some control of the policymaker, a policy-relevant parameter can describe the effects as a function of the willingness to participate represented by the propensity score. Then the following definition of the DMTE makes sense:

\[
DMTE(w, x, z) := E[\beta_i | P(X_i, Z_i) = P(x, z), W_i = w, X_i = x] ,
\]

From this last definition we can identify a policy-relevant parameter closer to the LATE than the DLATE as follows:

\[
E[\beta_i | P(x, z_2) \geq P(X_i, Z_i) > P(x, z_1), W_i = w, X_i = x] = \frac{1}{P(x, z_2) - P(x, z_1)} \int_{P(x, z_1)}^{P(x, z_2)} DMTE(u, w, x) du .
\]

However, this parameter is less informative than the LATE regarding the policy relevance as they are based on a change in the willingness to participate induced by a shift of the instrument from \( z_1 \) to \( z_2 \) rather than the relating intrinsic participation. In other words, an individual \( i \) with a probability to participate moving from \( P(X_i, z_1) \) to \( P(X_i, z_2) \) is more likely to participate without necessarily participating.

\[
DMTE(w, x, z) = \beta(P(x, z), x, w)
\]

or

\[
DMTE(u, w, x) = \beta(P(x, z) = u, x, w)
\]

The identification is similar to the continuous case. However, we consider the well known common support assumption that enables us to circumvent the large
interpolation issue mentioned in the continuous case. Thus, we can also consider
a kernel estimation in the present context.

Similarly to the continuous case, we can identified 
\( \beta(P(x, z) - d, x, w) \). Then, \( d \) should be integrated out to obtain \( DMTE(w, x, z) \). This step can be simplified by assuming the usual common support assumption as follows:

\[
\forall x, z, \quad 0 < P(x, z) < 1.
\]

Then, by marginal integration,

\[
\beta(P(x, z), x, w) = \sum_{d=0}^{1} \beta(P(x, z) - d, x, w) P(D = d|x, Z = z) \\
= \beta(P(x, z) - 1, x, w) P(x, z) + \beta(P(x, z), x, w) [1 - P(x, z)], \\
\Rightarrow \quad [\beta(P(x, z) - 1, x, w) - \beta(P(x, z), x, w)] P(x, z) = 0.
\]

Then, from the common support assumption, we have:

\[
\beta(P(x, z) - 1, x, w) = \beta(P(x, z), x, w).
\]

Similarly, we can show that

\[
\alpha(P(x, z) - 1, x, w) = \alpha(P(x, z), x, w).
\]

First, we can identify \( \alpha(P(x, z), x, w) \) as follows

\[
E[Y|V = P(x, z), W = w, X = x] = \alpha(P(x, z), x, w).
\]

Then,

\[
\beta(P(x, z) - 1, x, w) = \pi(P(x, z) - 1, x, w) - \alpha(P(x, z) - 1, x, w), \\
= \pi(P(x, z) - 1, x, w) - \alpha(P(x, z), x, w).
\]

Finally, we can identify,

\[
DMTE(w, x, z) = \beta(P(x, z), x, w) = \beta(P(x, z) - 1, x, w),
\]
as well as,

\[ DMTE(u, w, x) = \beta(P(x, z) = u, x, w) \, . \]

Now let us investigate the estimations of these parameters. The first stage concerns the estimation of \( P(x, z) \). If the dimension of \( x \) and \( z \) is small, then we can implement a logistic regression by the local maximum likelihood method obtained by kernel function weighting. However, in most cases, the dimension of \( X \) is too high to sustain a full non-parametric estimation. An intermediate alternative to mitigate the curse of dimensionality would be to assume additive splines of which the coefficients can be estimated by penalized maximum likelihood method with a binomial link function.

Unlike the continuous treatment case, large interpolation are not required thanks to the monotonicity assumptions. As a result, we propose here a simple kernel estimation based on the assumption that the dimension of \( x \) and \( w \) is small. In particular, a local polynomial is considered as it is coherent with the identification strategy provided in the previous section and can be easily implemented by the well known Weighted Least Square (WLS) method where weights are provided by a kernel function. The latter is parametrized by a bandwidth that considers and weights only the observations at the neighborhood of the interpolated point of the non-parametric function support.

From the multi-step identification considered in the previous section, we first estimate \( \alpha(P(x, z), w, x) \) of Model (3.12) by locally fitting a polynomial by WLS method as follows:

\[
\hat{\alpha}(\hat{P}(x, z), w, x) = \hat{E}[Y_i|\widehat{V}_i = \hat{P}(x, z), W_i = w, X_i = x] = \hat{\alpha}_0 \, .
\]

such that,

\[
(\hat{\alpha}_0, \hat{\alpha}_p) = \arg\min_{\alpha_0,\alpha_p} \sum_{i=1}^{n} [Y_i - \alpha_0 - P_d[\widehat{V}_i - \hat{P}(x, z), X_i - x, W_i - w]^T \alpha_d]^2 \\
\times K_h(\widehat{V}_i - \hat{P}(x, z))K_h(X_i - x)K_h(W_i - w) \, ,
\]

94
where $\mathcal{P}_d$ is the operator that yields the polynomial basis of order $d$ up to a constant term. Therefore, $\alpha_0 + \mathcal{P}_d[\hat{V}_i - \hat{P}(x, z), X_i - x, W_i - w]^T \alpha_d$ is the polynomial of order $d$ itself. $K_h$ is a kernel function that weights the observed values that are in the neighborhood. The latter is parametrized by the bandwidth $h$ that is often obtained by cross-validation. From Equation (3.14) we have,

\[
\hat{\text{DMTE}}(w, x, z) = \hat{\beta}(\hat{P}(x, z), w, x),
\]

\[
= E[Y_i | \hat{V}_i = \hat{P}(x, z) - 1, w, x] - \hat{\alpha}(\hat{P}(x, z), x, w),
\]

\[
= \hat{E} \left[ Y_i - \alpha_0 \hat{V}_i = \hat{P}(x, z) - 1, W_i = x, X_i = z \right],
\]

\[
= \hat{\beta}_0 ,
\]

such that,

\[
(\hat{\beta}_0, \hat{\beta}_d) = \arg \min_{\beta_0, \beta_d} \sum_{i=1}^{n} \left[ Y_i - \alpha_0 - \beta_0 - \mathcal{P}_d[\hat{V}_i - \hat{P}(x, z) + 1, X_i - x, W_i - x]^T \beta_d \right]^2 \times K_h(\hat{V}_i - \hat{P}(x, z) + 1) K_h(X_i - x) K_h(W_i - w).
\]

If the dimension of $(x, w)$ is too high, then a full non-parametric specification of $\alpha$ and $\beta$ is subject to the curse of dimensionality that can harm the reliability of the estimation. In this case, an additive structure of $\alpha$ and $\beta$ can be considered by wisely removing interactions up to the first equations specified in Specifications (3.15). Similarly, if $x$ and $w$ have themselves a high dimension, the additive structure can be extended up to the second equations of Specifications (3.15).

\[
\alpha(P(x, z) - d, x, w) := \alpha_0 + \alpha_v(P(x, z) - d) + \alpha_w(w) + \alpha_x(x) ,
\]

\[
:= \alpha_0 + \alpha_v(P(x, z) - d) + \sum_{l=1}^{L} \alpha_l(w_l) + \sum_{k=1}^{K} \alpha_k(x_k) ,
\]

\[
\beta(P(x, z) - d, x, w) := \beta_0 + \beta_v(P(x, z) - d) + \beta_w(w) + \beta_x(x) ,
\]

\[
:= \beta_0 + \beta_v(P(x, z) - d) + \sum_{l=1}^{L} \beta_l(w_l) + \sum_{k=1}^{K} \beta_k(x_k) .
\]

(3.15)
The kernel method presented above can be supplemented with a backfitting algorithm to estimate the additive structure presented in specifications (3.15). However, the backfitting algorithms for kernel estimation are seldom implemented by the statistical software. As a result, we also recommend to consider the spline methods proposes in the continuous treatment case.

Regarding the implementation, the same methods proposed in the previous section can be used. However, the binary treatment requires a different bootstrap strategy to build the confidence intervals. As a result, we consider the following algorithm:

1. For \( r = 1, ..., rep \) and \( i = 1, ..., n \).
2. Predict \( \hat{P}(X_i, Z_i) \) after regressing \( S \) on \( (X, Z) \).
3. Generate \( \hat{V}_i^r = \begin{cases} \hat{P}(X_i, Z_i) & \text{if } \hat{P}(X_i, Z_i) \leq U_i^r \\ \hat{P}(X_i, Z_i) - 1 & \text{otherwise} \end{cases} \) with \( U_i^r \sim U[0, 1] \).
4. Generate \( D_i^r = \hat{P}(X_i, Z_i) - \hat{V}_i^r \).
5. Predict \( \hat{P}^r(X_i, Z_i) \) after regressing \( D^r \) on \( (X, Z) \).
6. Predict \( \hat{\alpha}(\hat{V}_i, X_i, W_i) \) and \( \hat{\beta}(\hat{V}_i, X_i, W_i) \) after regressing \( Y \) on \( (\hat{V}_i, X_i, W_i) \).
7. Generate \( \hat{\sigma}_i = Y_i - \hat{\alpha}(\hat{V}_i, X_i, W_i) - \hat{\beta}(\hat{V}_i, X_i, W_i)D_i \).
8. Generate \( \varepsilon_i^r = \hat{\sigma}_i\Phi^{-1}(U_i^r) \).
9. Generate \( Y_i^r = \hat{\alpha}(\hat{P}^r(X_i, Z_i) - D_i^r, X_i, W_i) + \hat{\beta}(\hat{P}^r(X_i, Z_i) - D_i^r, X_i, W_i)D_i^r + \varepsilon_i^r \).
10. Predict \( \hat{\beta}'(\hat{P}^r(x, z), x, w) \) after regressing \( Y^r \) on \( (\hat{P}^r(X, Z) - D^r, D^r, X, W) \).

B. Proof of Proposition 3

Let \( i = 1, ..., n \) such that \( V_i = V(\cdot, X_i, Z_i) \) satisfies Model (3.1) endowed with the DAG depicted on Figure 3.1 with selection rule provided by Equation (3.3). Let
us assume that $V(., X_i, Z_i)$ is invertible.

Then, from the law of iterative expectations:

$$E[Y_i^d|V_i, W_i, X_i, Z_i] = E\{E[Y_i^d|A_i, V(A_i, X_i, Z_i), W_i, X_i, Z_i]|V_i, W_i, X_i, Z_i}\}
= E\{E[Y_i^d|A_i, W_i, X_i, Z_i]|V_i, W_i, X_i, Z_i\}$$

From the causal graph, we see that $Y_i^d \perp Z_i|(A_i, W_i, X_i)$. Therefore, $E[Y_i^d|A_i, W_i, X_i, Z_i] = E[Y_d|A_i, W_i, X_i]$ and then,

$$E[Y_i^d|V_i, W_i, X_i, Z_i] = E\{E[Y_i^d|A_i, W_i, X_i]|V_i, W_i, X_i, Z_i\}$$

Since $V(., X_i, Z_i)$ is invertible, we define $\lambda(V_i, X_i, Z_i) := V^{-1}(V_i, X_i, Z_i) = A_i$.

Then,

$$E[Y_i^d|V_i, W_i, X_i, Z_i] = E\{E[Y_i^d|\lambda(V_i, X_i, Z_i), W_i, X_i]|V_i, W_i, X_i, Z_i\}
= E[Y_i^d|A_i = \lambda(V_i, X_i, Z_i), W_i, X_i]
= E[Y_i^d|V_i = V(A_i, X_i, Z_i), W_i, X_i]
= E[Y_i^d|V_i, W_i, X_i]$$

C. Proof of Proposition 4

Let $i = 1, \ldots, n$ such that

$$Y_i = Y_i(0) + \lim_{J \to \infty} \sum_{j=1}^{J} [Y_i(j) - Y_i(j-1)] \mathbb{I}(S_i \geq j),$$

$$= Y_i(0) + \sum_{j=1}^{S_i} [Y_i(j) - Y_i(j-1)],$$

$S_i \in \mathbb{N}, \ i = 1, \ldots, n.$
The second term is coherent with the definition of the surface under a step function with unit length stairs. Therefore, a generalization to $S$ continuous will obviously yield to a Riemann integral. To see this, let us consider the continuous treatment with $S_i \in \mathbb{R}_+$. On this purpose, we define $b_i(t^*_j) := Y_i(j) - Y_i(j - 1)$ such that $t^*_j \in [j - 1, j[$ with $j = 1, ..., J$. Then we have

$$Y_i = Y_i(0) + \sum_{j=1}^{S_i} b_i(t^*_j),$$

$$= Y_i(0) + \sum_{j=1}^{S_i} b_i(t^*_j)(a_j - a_{j-1}),$$

where $a_j = j$ and $a_{j-1} = j - 1$. The second term is a Riemann sum with a unit interval $a_j - a_{j-1}$ for all $j$. To obtain the continuous treatment case, we consider the following limit:

$$Y_i = Y_i(0) + \lim_{a_{j-1} \rightarrow t^*_j \rightarrow a_j} \sum_{j=1}^{S_i} b_i(t^*_j)(a_j - a_{j-1}),$$

$$= Y_i(0) + \int_0^{S_i} b_i(t)dt,$$

$$= Y_i(0) + \left[B_i(t)\right]_0^{S_i},$$

such the second equality comes from the definition of the Riemann integral and $\int b_i(t)dt = B_i(t) + c$ with $c \in \mathbb{R}$. Then,

$$Y_i = Y^*_i + B_i(S_i),$$

where $B_i(.)$ is a random function such that $B_i(0) = 0$ a.s..
Bibliography


Appendices
Appendix A

Computer Script of Chapter 2

# # # # # # # # # # # # # # # # #
# Load Packages #
# # # # # # # # # # # # # # # # #

library(np)
library(mgcv)
library(gamlss)

# # # # # # # # # # # # # # # # #
# Design 1 #
# # # # # # # # # # # # # # # # #

n=250  # sample size
rep=500  # numbers of repetitions
predict_gam_g=numeric()
predict_np_g=numeric()
predict_gamlss_g=numeric()
predict_gam_h=numeric()
predict_np_h=numeric()
predict_gamlss_h=numeric()

# Generation of the data

z = seq(-2, 2, 4/(n-1))

for (i in 1:rep){

w1 = rnorm(n)
w2 = rnorm(n)

x1 = 2*w1 - 0.5*w2^2 + 2*z

g = cos(3*z)h = sin(2.5*z)
e=rnorm(n)

y <- g + x1*h + e

# Regressions with
\begin{verbatim}
m = gam(y ~ s(z) + s(z, by = x1))  #mgcv package
bw <- npscoefbw(y ~ x1 | z)  #np package
model.np <- npscoef(bw, betas=TRUE)
model.gamlss = gamlss(y ~ pvc(z) + pvc(z, by = x1))  #gamlss package

#Predictions of the varying-coefficients functions

p_gam = predict(m, type="terms")
p_gamlss = predict(model.gamlss, type="terms")

predict_gam_g = cbind(predict_gam_g, p_gam[,1])
predict_nop_g = cbind(predict_nop_g, coef(model.np)[,1])
predict_gamlss_g = cbind(predict_gamlss_g, p_gamlss[,1])

predict_gam_h = cbind(predict_gam_h, p_gam[,2]/x1)
predict_nop_h = cbind(predict_nop_h, coef(model.np)[,2])
predict_gamlss_h = cbind(predict_gamlss_h, p_gamlss[,2]/x1)

}

#Results

par(mfrow=c(3,2))

############ mgcv results ############

dim(predict_gam_g)
\end{verbatim}
interval_gamlss_g=apply(predict_gam_g,1,quantile,probs = c(.05, .95))
meanpredict_gam_g=apply(predict_gam_g,1,mean)

interval_gam_g=apply(predict_gam_g,1,quantile,probs = c(.05, .95))

plot(z,as.vector(meanpredict_gam_g),
     ylim=c(quantile(interval_gamlss_g[1,],.10),
     quantile(interval_gamlss_g[2,],.9)),
     main="MGCV",xlab="Z",ylab="g(.)")

lines(z,interval_gam_g[1,],type="l")
lines(z,interval_gam_g[2,],type="l")

curve(cos(3*x),add=TRUE,lty=2)

Bias_gam_g=mean(meanpredict_gam_g-g)
MSE_gam_g=mean((meanpredict_gam_g-g)^2)

dim(predict_gam_h)
interval_gam_h=apply(predict_gam_h,1,quantile,probs = c(.05, .95))
meanpredict_gam_h=apply(predict_gam_h,1,mean)

interval_gamlss_h=apply(predict_gamlss_h,1,quantile,probs = c(.05, .95))

plot(z,as.vector(meanpredict_gam_h),
     ylim=c(quantile(interval_gamlss_h[1,],.10),

108
quantile(interval_gamlss_h[2,], 0.9),
main="MGCV", xlab="Z", ylab="h(.)")

lines(z, interval_gaml_h[1,], type="l")
lines(z, interval_gaml_h[2,], type="l")

curve(sin(2.5*x), add=TRUE, lty=2)

Bias_gam_h=mean(meanpredict_gam_h-h)
MSE_gam_h=mean((meanpredict_gam_h-h)^2)

############################ np results ###############################

dim(predict_np_g)
interval_np_g=apply(predict_np_g, 1, quantile, probs = c(.05, .95))
meanpredict_np_g=apply(predict_np_g, 1, mean)

plot(z, as.vector(meanpredict_np_g),
     ylim=c(quantile(interval_gamlss_g[1,], 0.10),
            quantile(interval_gamlss_g[2,], 0.9)),
     main="NP", xlab="Z", ylab="g(.)")

lines(z, interval_np_g[1,], type="l")
lines(z, interval_np_g[2,], type="l")

curve(cos(3*x), add=TRUE, lty=2)

Bias_np_g=mean(meanpredict_np_g-g)
MSE_{np, g} = \text{mean}((\text{meanpredict}_{np, g} - g)^2)

\text{dim}(\text{predict}_{np, h})
\text{interval}_{np, h} = \text{apply}(\text{predict}_{np, h}, 1, \text{quantile}, \text{probs} = c(0.05, 0.95))
\text{meanpredict}_{np, h} = \text{apply}(\text{predict}_{np, h}, 1, \text{mean})

\text{plot}(z, \text{as.vector}(\text{meanpredict}_{np, h}),
\text{ylim}=c(\text{quantile}(\text{interval}_{gamlss}_{h}[1,], 0.10),
\text{quantile}(\text{interval}_{gamlss}_{h}[2,], 0.9)),
\text{main}="\text{NP}" , \text{xlab}="Z" , \text{ylab}="h(.)")

\text{lines}(z, \text{interval}_{np, h}[1,], \text{type}="1")
\text{lines}(z, \text{interval}_{np, h}[2,], \text{type}="1")
\text{curve}(\sin(2.5 \times x), \text{add}=\text{TRUE}, \text{lty}=2)

\text{Bias}_{np, h} = \text{mean}(\text{meanpredict}_{np, h} - h)
\text{MSE}_{np, h} = \text{mean}((\text{meanpredict}_{np, h} - h)^2)

######################################## gamlss results #######################################

\text{dim}(\text{predict}_{gamlss, g})
\text{interval}_{gamlss, g} = \text{apply}(\text{predict}_{gamlss, g}, 1, \text{quantile}, \text{probs} = c(0.05, 0.95))
\text{meanpredict}_{gamlss, g} = \text{apply}(\text{predict}_{gamlss, g}, 1, \text{mean})
plot(z, as.vector(meanpredict_gamlss_g),
    ylim=c(quantile(interval_gamlss_g[1,], 0.10),
    quantile(interval_gamlss_g[2,], 0.9)),
    main="GAMLSS", xlab="Z", ylab="g(.)")

lines(z, interval_gamlss_g[1,], type="l")
lines(z, interval_gamlss_g[2,], type="l")

curve(cos(3*x), add=TRUE, lty=2)

Bias_gamlss_g=mean(meanpredict_gamlss_g-g)
MSE_gamlss_g=mean((meanpredict_gamlss_g-g)^2)

dim(predict_gamlss_h)
interval_gamlss_h=apply(predict_gamlss_h, 1, quantile, probs = c(0.05, 0.95))
meanpredict_gamlss_h=apply(predict_gamlss_h, 1, mean)

plot(z, as.vector(meanpredict_gamlss_h),
    ylim=c(quantile(interval_gamlss_h[1,], 0.10),
    quantile(interval_gamlss_h[2,], 0.9)),
    main="GAMLSS", xlab="Z", ylab="h(.)")

lines(z, interval_gamlss_h[1,], type="l")
lines(z, interval_gamlss_h[2,], type="l")

111
curve(sin(2.5*x),add=TRUE,lty=2)

Bias_gamlss_h=mean(meanpredict_gamlss_h−h)
MSE_gamlss_h=mean((meanpredict_gamlss_h−h)^2)

######################## Table of Results ########################

cbind(Bias_gam_g,Bias_gam_h,MSE_gam_g,MSE_gam_h)
cbind(Bias_np_g,Bias_np_h,MSE_np_g,MSE_np_h)
cbind(Bias_gamlss_g,Bias_gamlss_h,MSE_gamlss_g,MSE_gamlss_h)

####################
# Design 2 #
####################

predict_gam_g=numeric()
predict_gamlss_g=numeric()
predict_gam_h=numeric()
predict_gamlss_h=numeric()

# Generation of the data
for (i in 1:rep){
    x1 = runif(n,−2,2)
    x2 = 0.25*x1+0.75*runif(n,−2,2)
\begin{verbatim}
g = \cos(3*x1) 
h = \sin(2.5*x2) 
e = \text{rnorm}(n) 
y <- 6 + x1*g + x2*h + e 

# Regressions with 
m = \text{gam}(y \sim s(x1, \text{by} = x1) + s(x2, \text{by} = x2)) 
# mgcv package 
model_gamlss = \text{gamlss}(y \sim \text{pvc}(x1, \text{by} = x1) + \text{pvc}(x2, \text{by} = x2)) 
# gamlss package 

# Predictions of Varying-coefficient functions 

p_gam = \text{predict}(m, \text{type}="terms") 
p_gamlss = \text{predict}(model_gamlss, \text{type}="terms") 

data = \text{data.frame}(x1, x2) 
newdata = \text{data.frame}(x1, x2) 
newdata$x1 = \text{seq}(-2, 2, 4/(n-1)) 
newdata$x2 = \text{seq}(-2, 2, 4/(n-1)) 

p_gam = \text{predict}(m, \text{type}="terms", \text{data}=data, \text{newdata}=newdata) 
p_gamlss = \text{predict}(model_gamlss, \text{type}="terms", \text{data}=data, \text{newdata}=newdata) 

\text{predict.gam.g} = \text{cbind}(\text{predict.gam.g}, p_gam[,1]/newdata$x1) 
\end{verbatim}
predict_gamlss_g=cbind(predict_gamlss_g, p_gamlss[,1]/newdata$x1)

predict_gam_h=cbind(predict_gam_h, p_gam[,2]/newdata$x2)
predict_gamlss_h=cbind(predict_gamlss_h, p_gamlss[,2]/newdata$x2)

# Results

par(mfrow=c(2,2))

# mgcv results

dim(predict_gam_g)
interval_gam_g=apply(predict_gam_g,1,quantile,probs = c(.05, .95))
meanpredict_gam_g=apply(predict_gam_g,1,mean)

interval_gamlss_g=apply(predict_gamlss_g,1,quantile,probs = c(.05, .95))

plot(newdata$x1, as.vector(meanpredict_gam_g),
ylim=c(quantile(interval_gamlss_g[1,],0.10),
quantile(interval_gamlss_g[2,],0.9)),
main="MCCV", xlab="X_1", ylab="g(.)"
)

lines(newdata$x1, interval_gam_g[1,], type="1")
lines(newdata$x1, interval_gam_g[2,], type="1")

curve(cos(3*x), add=TRUE, lty=2)
Bias_gam_g = mean(meanpredict_gam_g - cos(3*newdata$x1))
MSE_gam_g = mean((meanpredict_gam_g - cos(3*newdata$x1))^2)

dim(predict_gam_h)
interval_gam_h = apply(predict_gam_h, 1, quantile, probs = c(0.05, 0.95))
meanpredict_gam_h = apply(predict_gam_h, 1, mean)

interval_gamlss_h = apply(predict_gamlss_h, 1, quantile, probs = c(0.05, 0.95))

plot(newdata$x2, as.vector(meanpredict_gam_h),
     ylim = c(quantile(interval_gamlss_h[1,], 0.10),
              quantile(interval_gamlss_h[2,], 0.9)),
     main = "MGCV", xlab = "X_2", ylab = "h(.)")

lines(newdata$x2, interval_gam_h[1,], type = "l")
lines(newdata$x2, interval_gam_h[2,], type = "l")

curve(sin(2.5*x), add = TRUE, lty = 2)

Bias_gam_h = mean(meanpredict_gam_h - sin(2.5*newdata$x2))
MSE_gam_h = mean((meanpredict_gam_h - sin(2.5*newdata$x2))^2)

############################ gamlss results ###################################

115
\texttt{dim(predict\_gamlss\_g)}
\texttt{interval\_gamlss\_g=apply(predict\_gamlss\_g,1,quantile, probs = c(.05, .95))}
\texttt{meanpredict\_gamlss\_g=apply(predict\_gamlss\_g,1,mean)}
\texttt{plot(newdata$x1, as.vector(meanpredict\_gamlss\_g),
ylim=c(quantile(interval\_gamlss\_g[1,],0.10),
quantile(interval\_gamlss\_g[2,],0.9)),
main="GAMLSS" , xlab="X\_1" , ylab="g(.)")}
\texttt{lines(newdata$x1, interval\_gamlss\_g[1,], type="l")
lines(newdata$x1, interval\_gamlss\_g[2,], type="l")
}
\texttt{curve(cos(3*x), add=TRUE, lty =2)}
\texttt{Bias\_gamlss\_g=mean(meanpredict\_gamlss\_g-cos(3*newdata$x1))}
\texttt{MSE\_gamlss\_g=mean((meanpredict\_gamlss\_g-cos(3*newdata$x1))^2)}
\texttt{dim(predict\_gamlss\_h)}
\texttt{interval\_gamlss\_h=apply(predict\_gamlss\_h,1,quantile, probs = c(.05, .95))}
\texttt{meanpredict\_gamlss\_h=apply(predict\_gamlss\_h,1,mean)}
\texttt{plot(newdata$x2, as.vector(meanpredict\_gamlss\_h),
ylim=c(quantile(interval\_gamlss\_h[1,],0.10),
quantile(interval\_gamlss\_h[2,],0.9)),
main="GAMLSS" , xlab="X\_2" , ylab="h(.)")}
lines(newdata$x2, interval_gamlss_h[1,], type="l")
lines(newdata$x2, interval_gamlss_h[2,], type="l")

curve(sin(2.5*x), add=TRUE, lty=2)

Bias_gamlss_h=mean(meanpredict_gamlss_h-sin(2.5*newdata$x2))
MSE_gamlss_h=mean((meanpredict_gamlss_h-sin(2.5*newdata$x2))^2)

############################ Table of results ##################################

cbind(Bias_gam_g, Bias_gam_h, MSE_gam_g, MSE_gam_h)
cbind(Bias_gamlss_g, Bias_gamlss_h, MSE_gamlss_g, MSE_gamlss_h)

############################
# Design 3 #
############################

rm(list=ls())
n=250
rep=500
predict_gam_g=numeric()
predict_gamlss_g=numeric()
predict_gam_h=numeric()
predict_gamlss_h=numeric()

# Generation of the Data
for (i in 1:rep) {

z1 = runif(n, -2, 2)
z2 = 0.25*z1 + 0.75*runif(n, -2, 2)

w1 = rnorm(n)
w2 = rnorm(n)

x1 = 0.25*w1 + 0.4*w2 + z2
x2 = 2*w1 - 0.5*w2^2 + 2*z1


g= cos(3*z1)
h= sin(2.5*z2)

e=rnorm(n)

 y <- 6 + x1*g + x2*h + e

# Regressions with

m=gam(y~s(z1, by = x1)+s(z2, by = x2))
# mgcv package
model_gamlss = gamlss(y~pvc(z1, by = x1)+pvc(z2, by = x2))
# gamlss package

# Predictions of Varying-coefficient functions
\texttt{data=data.frame(z1, z2, x1, x2)}
\texttt{newdata=data.frame(z1, z2, x1, x2)}
\texttt{newdata$z1=seq(-2,2,4/(n-1))}
\texttt{newdata$z2=seq(-2,2,4/(n-1))}

\texttt{p_gam=predict(m, type="terms", data=data, newdata=newdata)}
\texttt{p_gamlss=predict(model_gamlss, type="terms", data=data, newdata=newdata)}

\texttt{predict_gam_g=cbind(predict_gam_g, p_gam[,1]/newdata$x1)}
\texttt{predict_gamlss_g=cbind(predict_gamlss_g, p_gamlss[,1]/newdata$x1)}

\texttt{predict_gam_h=cbind(predict_gam_h, p_gam[,2]/newdata$x2)}
\texttt{predict_gamlss_h=cbind(predict_gamlss_h, p_gamlss[,2]/newdata$x2)}

\}

\# Results

\texttt{par(mfrow=c(2,2))}

\### mgcv results

\texttt{dim(predict_gam_g)}
\texttt{interval_gam_g=apply(predict_gam_g, 1, \texttt{quantile}, probs = c(0.05, 0.95))}
\texttt{meanpredict_gam_g=apply(predict_gam_g, 1, \texttt{mean})}

\texttt{interval_gamlss_g=apply(predict_gamlss_g, 1, \texttt{quantile}, probs = c(0.05, 0.95))}
plot(newdata$z1, as.vector(meanpredict_gam_g),
    ylim=c(quantile(interval_gamlss_g[1,], 0.10),
    quantile(interval_gamlss_g[2,], 0.9)),
    main="M G C V", xlab="Z_1", ylab="g(.)")
lines(newdata$z1, interval_gam_g[1,], type="l")
lines(newdata$z1, interval_gam_g[2,], type="l")

curve(cos(3*x), add=TRUE, lty=2)

Bias_gam_g=mean(meanpredict_gam_g-cos(3*newdata$z1))
MSE_gam_g=mean((meanpredict_gam_g-cos(3*newdata$z1))^2)

dim(predict_gam_h)
interval_gam_h=apply(predict_gam_h, 1, quantile, probs = c(0.05, 0.95))
meanpredict_gam_h=apply(predict_gam_h, 1, mean)

interval_gamlss_h=apply(predict_gamlss_h, 1, quantile, probs = c(0.05, 0.95))

plot(newdata$z2, as.vector(meanpredict_gam_h),
    ylim=c(quantile(interval_gamlss_h[1,], 0.10),
    quantile(interval_gamlss_h[2,], 0.9)),
    main="M G C V", xlab="Z_2", ylab="h(.)")
lines(newdata$z2, interval_gam_h[1,], type="l")
lines(newdata$z2, interval_gam_h[2,], type="1")

curve(sin(2.5*x), add=TRUE, lty=2)

Bias_gam_h=mean(meanpredict_gam_h−sin(2.5*newdata$z2))
MSE_gam_h=mean((meanpredict_gam_h−sin(2.5*newdata$z2))^2)

# # # # # # # # # # # # # # gamlss results # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #

dim(predict_gamlss_g)
interval_gamlss_g=apply(predict_gamlss_g, 1, quantile, probs = c(.05, .95))
meanpredict_gamlss_g=apply(predict_gamlss_g, 1, mean)

plot(newdata$z1, as.vector(meanpredict_gamlss_g),
     ylim=c(quantile(interval_gamlss_g[1,], 0.10),
            quantile(interval_gamlss_g[2,], 0.9)),
     main="GAMLSS", xlab="Z_1", ylab="g(.)")

lines(newdata$z1, interval_gamlss_g[1,], type="1")
lines(newdata$z1, interval_gamlss_g[2,], type="1")

curve(cos(3*x), add=TRUE, lty=2)

Bias_gamlss_g=mean(meanpredict_gamlss_g−cos(3*newdata$z1))
MSE_gamlss_g=mean((meanpredict_gamlss_g−cos(3*newdata$z1))^2)
dim(predict_gamlss_h)
interval_gamlss_h=apply(predict_gamlss_h,1,quantile,probs = c(.05, .95))
meanpredict_gamlss_h=apply(predict_gamlss_h,1,mean)

plot(newdata$z2, as.vector(meanpredict_gamlss_h),
      ylim=c(quantile(interval_gamlss_h[1,],0.10),
                                  quantile(interval_gamlss_h[2,],0.9)),
      main="GAMLSS", xlab="Z_2", ylab="h(.)")
lines(newdata$z2, interval_gamlss_h[1,], type="l")
lines(newdata$z2, interval_gamlss_h[2,], type="l")

curve(sin(2.5*x),add=TRUE, lty=2)

Bias_gamlss_h=mean(meanpredict_gamlss_h-sin(2.5*newdata$z2))
MSE_gamlss_h=mean((meanpredict_gamlss_h-sin(2.5*newdata$z2))^2)

#########################################Table of Results#########################################

cbind(Bias_gam_g, Bias_gam_h,MSE_gam_g,MSE_gam_h)
cbind(Bias_gamlss_g, Bias_gamlss_h,MSE_gamlss_g,MSE_gamlss_h)

#########################################
# Design 4 #
#########################################
# Generation of the data

for (i in 1:rep){

z1 = runif(n,-2,2)
v = runif(n,-2,2)

w1 = rnorm(n)
w2 = rnorm(n)

x1 = 0.25*z1 + 0.4*w2 + v

g= cos(3*z1)
h= sin(2.5*v)
u = h + rnorm(n)

e=rnorm(n)

y <- 6+x1*g + x1*u + e
# First stage regression

\[
\text{vhat} = \text{resid}(\text{gam}(x_1 \sim s(z_1, w_2)))
\]

# Second stage regressions with

\[
\text{m} = \text{gam}(y \sim s(z_1, \text{by} = x_1) + s(\text{vhat}, \text{by} = x_1))
\]

# mgcv package

\[
\text{model_gamlss} = \text{gamlss}(y \sim \text{pvc}(z_1, \text{by} = x_1) + \text{pvc}(\text{vhat}, \text{by} = x_1))
\]

# gamlss package

# Predictions of the varying-coefficient functions

\[
\text{data} = \text{data.frame}(z_1, \text{vhat}, x_1)
\]

\[
\text{newdata} = \text{data.frame}(z_1, \text{vhat}, x_1)
\]

\[
\text{newdata}\$z1 = \text{seq}(-2, 2, 4/(n-1))
\]

\[
\text{newdata}\$\text{vhat} = \text{seq}(-2, 2, 4/(n-1))
\]

\[
\text{p_gam} = \text{predict}(\text{m}, \text{type} = \text{"terms"}, \text{data} = \text{data}, \text{newdata} = \text{newdata})
\]

\[
\text{p_gamlss} = \text{predict}(\text{model_gamlss}, \text{type} = \text{"terms"}, \text{data} = \text{data}, \text{newdata} = \text{newdata})
\]

\[
\text{predict_gam_g} = \text{cbind}(\text{predict_gam_g},
\text{p_gam}[ , 1]/\text{newdata}\$x1 - \text{mean}(\text{p_gam}[ , 1]/\text{newdata}\$x1))
\]

\[
\text{predict_gamlss_g} = \text{cbind}(\text{predict_gamlss_g}, \text{p_gamlss}[, 1]/\text{newdata}\$x1)
\]

\[
\text{predict_gam_h} = \text{cbind}(\text{predict_gam_h},
\text{p_gam}[ , 2]/\text{newdata}\$x1 - \text{mean}(\text{p_gam}[ , 2]/\text{newdata}\$x1))
\]

\[
\text{predict_gamlss_h} = \text{cbind}(\text{predict_gamlss_h}, \text{p_gamlss}[ , 2]/\text{newdata}\$x1)
\]
# Results

```r
par(mfrow=c(2,2))

############### mgcv results ###############

dim(predict_gam_g)
interval_gam_g=apply(predict_gam_g, 1, quantile, probs = c(.05, .95))
meanpredict_gam_g=apply(predict_gam_g, 1, mean)

interval_gamlss_g=apply(predict_gamlss_g, 1, quantile, probs = c(.05, .95))

plot(newdata$z1, as.vector(meanpredict_gam_g),
     ylim=c(quantile(interval_gamlss_g[1,], 0.10),
           quantile(interval_gamlss_g[2,], 0.9)),
     main="MGCV", xlab="Z", ylab="g(.)")

lines(newdata$z1, interval_gam_g[1,], type="l")
lines(newdata$z1, interval_gam_g[2,], type="l")

curve(cos(3*x), add=TRUE, lty=2)

Bias_gam_g=mean(meanpredict_gam_g-cos(3*newdata$z1))
MSE_gam_g=mean((meanpredict_gam_g-cos(3*newdata$z1))^2)
```

125
\texttt{dim(\texttt{predict\_gam\_h})
interval\_gam\_h=apply(\texttt{predict\_gam\_h,1,quantile},probs = c(.05,.95))
meanpredict\_gam\_h=apply(\texttt{predict\_gam\_h,1,mean})

interval\_gamlss\_h=apply(\texttt{predict\_gamlss\_h,1,quantile},probs = c(.05,.95))

\texttt{plot(newdata$\texttt{vhat,as\.vector(meanpredict\_gam\_h),
    ylim=c(quantile(interval\_gamlss\_h[1,],0.10),
    quantile(interval\_gamlss\_h[2,],0.9)),
    main="MGCV",xlab="v",ylab="h(.)")

lines(newdata$\texttt{vhat,interval\_gam\_h[1,],type="l"})
lines(newdata$\texttt{vhat,interval\_gam\_h[2,],type="l"})

curve(sin(2.5*x),add=TRUE,lty=2)

Bias\_gam\_h=mean(meanpredict\_gam\_h−sin(2.5*newdata$\texttt{v}))
MSE\_gam\_h=mean((meanpredict\_gam\_h−sin(2.5*newdata$\texttt{v}))^2)

############################ gamlss results ####################################
plot(newdata$z1, as.vector(meanpredict_gamlss_g),
    ylim=c(quantile(interval_gamlss_g[1,], 0.10),
    quantile(interval_gamlss_g[2,], 0.9)),
    main="GAMLSS",
    xlab="Z", ylab="g(.)")

lines(newdata$z1, interval_gamlss_g[1,], type="l")
lines(newdata$z1, interval_gamlss_g[2,], type="l")

curve(cos(3*x), add=TRUE, lty=2)

Bias_gamlss_g = mean(meanpredict_gamlss_g - cos(3*newdata$z1))
MSE_gamlss_g = mean((meanpredict_gamlss_g - cos(3*newdata$z1))^2)

dim(predict_gamlss_h)
interval_gamlss_h = apply(predict_gamlss_h, 1, quantile, probs = c(0.05, 0.95))
meanpredict_gamlss_h = apply(predict_gamlss_h, 1, mean)

plot(newdata$vhat, as.vector(meanpredict_gamlss_h),
    ylim=c(quantile(interval_gamlss_h[1,], 0.10),
    quantile(interval_gamlss_h[2,], 0.9)),
    main="GAMLSS", xlab="v", ylab="h(.)")

lines(newdata$vhat, interval_gamlss_h[1,], type="l")
lines(newdata$vhat, interval_gamlss_h[2,], type="l")
curve(sin(2.5*x),add=TRUE, lty=2)

Bias_gamlss_h=mean(meanpredict_gamlss_h-sin(2.5*newdata$v))
MSE_gamlss_h=mean((meanpredict_gamlss_h-sin(2.5*newdata$v))^2)

######################## Table of results########################

cbind(Bias_gam_g,Bias_gam_h,MSE_gam_g,MSE_gam_h)
cbind(Bias_gamlss_g,Bias_gamlss_h,MSE_gamlss_g,MSE_gamlss_h)
List of Figures

1.1 $X^*$ is the consumption of Coca, while $Q^*$ is the consumption of Pepsi.  7

1.2 The (individual) demand functions for a given budget constraint

\[ p_1 X + p_2 Q = I \]

varies accordingly to individual preferences. Agent

3 and 4 do not consider Coca-Cola and Pepsi to be equally good
(left graph), while Agent 1 and 2 do (right graph). . . . . . . . . . . 8

1.3 A causal graph of (1.6) highlighting the causal links among the

variables. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14

1.4 The mechanism of the endogeneity process changes depending on

the assumptions about the relationship between the error $\varepsilon$ and the

stochastic disturbances $(\eta, \delta)$. The left picture is the only possibility

in a world of homogeneous coefficients, while the right specification

(with $cov(\eta, \delta) \neq 0$) is the situation resulting from introducing a

varying coefficient structure. The direct connection between $\delta$ and

$\varepsilon$ is not taken into account because the interest is about the causal

link between $Y$ and $W$ for a given level of $F$. . . . . . . . . . . . . 17

1.5 When endogeneity is an issue due to the presence of a model mis-

specification, the use of VCM is not enough to guarantee causal

analysis, and the introduction of IVs becomes necessary to ensure

the exogeneity. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24
2.1 Design 1, 90% simulation intervals for sample size $n = 50$ for $g$ on the left, $h$ on the right. .................................................. 49
2.2 Design 1, 90% simulation intervals for sample size $n = 100$ for $g$ on the left, $h$ on the right. .................................................. 50
2.3 Design 1, 90% simulation intervals for sample size $n = 250$ for $g$ on the left, $h$ on the right. .................................................. 51
2.4 Design 2, 90% simulation intervals for sample size $n = 50$ for $g$ on the left, $h$ on the right. .................................................. 52
2.5 Design 2, 90% simulation intervals for sample size $n = 100$ for $g$ on the left, $h$ on the right. .................................................. 53
2.6 Design 2, 90% simulation intervals for sample size $n = 250$ for $g$ on the left, $h$ on the right. .................................................. 54
2.7 Design 3, 90% simulation intervals for sample size $n = 50$ for $g$ on the left, $h$ on the right. .................................................. 55
2.8 Design 3, 90% simulation intervals for sample size $n = 100$ for $g$ on the left, $h$ on the right. .................................................. 56
2.9 Design 3, 90% simulation intervals for sample size $n = 250$ for $g$ on the left, $h$ on the right. .................................................. 57
2.10 Design 4, 90% simulation intervals for sample size $n = 50$ for $g$ on the left, $h$ on the right. .................................................. 58
2.11 Design 4, 90% simulation intervals for sample size $n = 100$ for $g$ on the left, $h$ on the right. .................................................. 59
2.12 Design 4, 90% simulation intervals for sample size $n = 250$ for $g$ on the left, $h$ on the right. .................................................. 60
3.1 DAG representation of the Generalized Roy Potential Outcome ($Y^d$) Model with unobserved abilities ($A$) correlated with the instruments $Z$ .............................................................. 72

3.2 Predicted Non-parametric both intercept $Y^0$ and slope DMTE as functions of $z$ in Model (3.12) .................................................. 85

3.3 Histogram of the instrument $Z :=$ SIMMED ........................................ 86
List of Tables

2.1 Simulated Bias over the Correlation between W and X . . . . . . . . 37
2.2 Results calculated from 500 samples drawn from Design 1. . . . . . 49
2.3 Results calculated from samples drawn from Design 2. . . . . . . . 51
2.4 Results calculated from 500 samples drawn from Design 3. . . . . . 54
2.5 Results calculated from 500 samples drawn from Design 4. . . . . . 56