FUNDAÇÃO GETULIO VARGAS
SCHOOL OF APPLIED MATHEMATICS

CAIO F. LINS PEIXOTO

NONPARAMETRIC INSTRUMENTAL VARIABLE REGRESSION THROUGH STOCHASTIC GRADIENTS AND KERNEL METHODS

Rio de Janeiro
2023

CAIO F. LINS PEIXOTO

NONPARAMETRIC INSTRUMENTAL VARIABLE REGRESSION THROUGH STOCHASTIC GRADIENTS AND KERNEL METHODS

Bachelor's dissertation presented to the School of Applied Mathematics (FGV/EMAp) to obtain the Bachelor's degree in Applied Mathematics.

Area of Study: Nonparametric Regression, Instrumental Variables, Kernel Methods, Stochastic Optimization, Machine Learning.

Advisor: Yuri F. Saporito

Rio de Janeiro

2023

Ficha catalográfica elaborada pela BMHS/FGV

Lins, Caio
Nonparametric Instrumental Variable Regression Through Stochastic Gradients and Kernel Methods/ Caio F. Lins Peixoto. – 2023.
26f.

Bachelor's Dissertation (Undergraduate) – School of Applied Mathematics.

Advisor: Yuri F. Saporito.
Includes bibliography.

1. Nonparametric Regression 2. Instrumental Variables 2. Stochastic Optimization I. Saporito, Yuri Fahham II. School of Applied Mathematics. III. Nonparametric Instrumental Variable Regression Through Stochastic Gradients and Kernel Methods

CAIO F. LINS PEIXOTO

NONPARAMETRIC INSTRUMENTAL VARIABLE REGRESSION THROUGH STOCHASTIC GRADIENTS AND KERNEL METHODS

Bachelor's dissertation presented to the School of Applied Mathematics (FGV/EMAp) to obtain the Bachelor's degree in Applied Mathematics.

Area of Study: Nonparametric Regression, Instrumental Variables, Kernel Methods, Stochastic Optimization, Machine Learning.

Approved on December —-, 2023
By the organizing committee

Yuri F. Saporito
School of Applied Mathematics

Board Member 1
Institution 1

Board Member 2
Institution 2

I dedicate this thesis to …

Acknowledgements

Thanks, ...

“ Bipedi bopedi bum! ”

Albert Einstein

Abstract

Keywords:

Resumo

Palavras-chave:

List of Figures

Figure 1 – Causal diagram for equation (2.1), where X is endogenous and Z is an IV. Source: prepared by the author. 15
Figure 2 – Comparison between OLS and 2SLS under two different levels of endogeneity. 19

List of Tables

Contents

1 INTRODUCTION 12

2 INSTRUMENTAL VARIABLE REGRESSION 13
2.1 Endogeneity 13
2.2 Instrumental Variables 14
2.3 Two Stages Least Squares (2SLS) 15
2.3.1 Constructing the estimator 16
2.3.2 Numerical examples 18
2.4 Nonparametric Instrumental Variable Regression 19
2.4.1 Problem specification 20
2.4.2 Identification 20

3 CONCLUSION 24

References 25

APPENDIX 26

1 Introduction

Remember to cite every person (NEWEY; POWELL, 2003).

2 Instrumental Variable Regression

This chapter provides an introduction to both parametric and nonparametric instrumental variable regression. It's goal is twofold. Firstly, we want to introduce the subject to readers unfamiliar with it. To make the exposition more fluid, we chose to delay the precise definition of all mathematical objects involved until Section 2.4, which deals with the nonparametric approach. The second goal is to precisely state the nonparametric regression problem which will be addressed in the remainder of this thesis. Along the exposition, we will cover the basics of Two Stages Least Squares, the IV regression method most widely employed in practice.

2.1 Endogeneity

We start by introducing the problem of endogenous covariates. The structural equation we consider is the following:

$$Y = h^\star(X) + \varepsilon, \tag{2.1}$$

where X is a vector of explanatory variables, Y is the scalar response, ε is a zero mean noise and the function $h^\star$ is the structural parameter we would like to estimate. The simplest estimation method for this model specification — and, therefore, one we would like to be able to use — is ordinary least squares (OLS), which works by finding, within a given class of functions $\mathcal{H}$, the element which minimizes the mean squared error:

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \mathbb{E}[(Y - h(X))^2]. \tag{2.2}$$

A reasonable and ample choice for $\mathcal{H}$ is the set of all square-integrable functions of X, that is, such that $\mathbb{E}[h(X)^2] < \infty$. Under this choice, we recover the conditional expectation of Y given X, i.e., $\hat{h}(X) = \mathbb{E}[Y \mid X]$. Expanding Y through (2.1), we find that $\hat{h}(X) = h^\star(X) + \mathbb{E}[\varepsilon \mid X]$. Hence, if $\mathbb{E}[\varepsilon \mid X]$ is not identically null, we have introduced bias in our estimation.

This is one of the problems which appear when $\mathbb{E}[\varepsilon \mid X] \neq 0$, or, more generally, when X and ε are correlated in some way. When this happens, we say that X is *endogenous*. There are several causes for endogenous covariates, the most common of which are (WOOLDRIDGE, 2001):

Omitted Variables This means ε can be decomposed as $g^\star(W)+\eta$, where $\mathbb{E}[\eta \mid X, W] = 0$ and X and W are correlated. Hence, when we don't observe W and leave it to the error

term, we end up estimating

$$\begin{aligned}\mathbb{E}[Y \mid X] &= h^\star(X) + \mathbb{E}[\varepsilon \mid X] \\ &= h^\star(X) + \mathbb{E}[g^\star(W) + \eta \mid X] \\ &= h^\star(X) + \mathbb{E}[g^\star(W) \mid X]\end{aligned}$$

which is likely different from $h^\star(X)$, if W is correlated with X. For example, if we want to regress a person's wage solely on her number of schooling years (this is X), there are other variables, unaccounted for, which influence both wages and schooling, such as natural ability (this is W). Innately skilled people may tend to be successful in school — and, therefore, pursue higher levels of education — as well as show higher performance in their future jobs, resulting in better wages. Thus, we fail to estimate $h^\star$.

Measurement Error If we are unable to exactly measure one of the covariates, X_k, and instead measure X_k' subject to some stochastic error, by using X_k' in our regression instead of X_k we are delegating to ε some measure of the difference between X_k and X_k'. Depending on how these two variables are related, we may introduce endogeneity. For example, X_k may be a marginal tax rate, but we may only have access to an average tax rate X_k'.

Simultaneity Simultaneity arises when one covariate X_k is determined simultaneously with Y. For example, if we are regressing neighborhood murder rates using the size of the local task force as a covariate, there is a simultaneity problem, since larger murder rates in a place cause a larger task force to be allocated there.

As we have said, bias in the estimation procedure is only one of the problems which arise when there are endogenous covariates. It's well known that the OLS estimate for linear regression fails to be consistent if any one of the covariates is endogenous (WOOLDRIDGE, 2001). To overcome endogeneity a few approaches exist, but by far the one most used by empirical economic research is instrumental variable estimation (WOOLDRIDGE, 2001).

2.2 Instrumental Variables

2.1 Definition An *instrumental variable* for regression problem (2.1) is a random variable Z such that

(i) There is some influence of Z upon X, that is, the marginal distribution of X is not the same as the distribution of X conditioned on Z;

(ii) The conditional mean of ε given Z is almost surely null, i.e., $\mathbb{E}[\varepsilon \mid Z] = 0$.

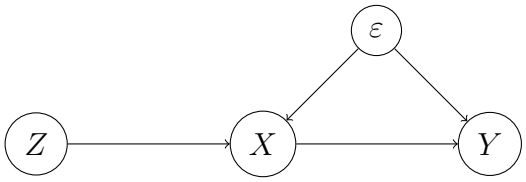

Figure 1 – Causal diagram for equation (2.1), where X is endogenous and Z is an IV. Source: prepared by the author.

The idea behind an instrumental variable is that it is exogenous (ii) while still influencing Y through X (i). An exogenous covariate, in contrast to an endogenous one, as a variable that is determined outside of the system described by (2.1).

Condition (ii) is only one of the possible meanings for the statement that Z is exogenous. Two possible alternatives are requiring that Z be (1) independent from, or (2) uncorrelated with ε. Of course, (1) is a much more strict requirement which implies (ii), while (2) is a softer condition, implied by (ii). Independence is almost always impossible to verify in real scenarios, so (1) is not a good option. In contrast, there are situations where condition (2) is enough for ensuring good properties of IV estimators, including one we will present shortly, the linear model (WOOLDRIDGE, 2001). However, in order to prepare grounds for the nonparametric methods that will come later, we chose to use the definition which serves both.

Instrumental variables are also studied in the context of causal inference, where the conditions above are presented differently, in terms of causal diagrams. In this field, instrumental variables are also required that to satisfy a third condition, phrased in terms of the causal diagram describing the relations between variables of interest (HERNÁN; ROBINS, 2020):

(iii) All paths from Z to Y must pass through X, that is, Z *only* influences Y through X.

In this sense, a typical causal diagram for an IV problem is the one in Figure 1.

2.3 Two Stages Least Squares (2SLS)

In this section, we restrict the structural function $h^\star$ in (2.1) to be affine:

$$h^\star(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{d_X} x_{d_X}, \tag{2.3}$$

and assume to have access to a random variable Z, taking values in $\mathbf{R}^{dz}$, satisfying conditions 2.1 (i) and (ii), so that Z is a valid instrumental variable. To ease the notation, we augment the variables X and Z to have a 1 as a first coordinate, so we may write $h^\star(X) = \beta^\top X$, where $\beta = (\beta_0, \ldots, \beta_{d_X})$. Their new dimensions are $d'_X = d_X + 1$ and

$d'_Z = d_Z + 1$. Our data is then composed of n independent joint samples $\{(X_i, Z_i, Y_i)\}_{i=1}^n$. Let $\boldsymbol{X} \in \mathbf{R}^{n \times d'_X}$ and $\boldsymbol{Z} \in \mathbf{R}^{n \times d'_Z}$ be the experiment design matrices with 1's in the first column, and let $\boldsymbol{Y} \in \mathbf{R}^n$ be the vector with all observations of Y. Each line of $\boldsymbol{X}$ and $\boldsymbol{Z}$ corresponds to one sample of the vectors X and Z, respectively. The idea of 2SLS is to first perform a regression of X on Z (the *first stage*) and then regress Y on the fitted values $\hat{X}$ (the *second stage*). In what follows, we will derive this method and give some numerical examples to show its applicability. To avoid misunderstandings during computations, we explicitly state that all vectors are regarded as *column* vectors.

2.3.1 Constructing the estimator

Since we have access to an exogenous covariate, a possible idea is to use this covariate to extract from X a component which is uncorrelated with ε. The simplest way to do this is to perform the *linear orthogonal projection* of X onto Z, that is, to find the matrix P which minimizes the MSE:

$$P = \underset{M \in \mathbf{R}^{d'_X \times d'_Z}}{\arg\min}\ \mathbb{E}[\|X - MZ\|^2] = \mathcal{L}(M).$$

A straightforward computation shows that $\nabla\mathcal{L}(M) = M\mathbb{E}[ZZ^\top] - \mathbb{E}[XZ^\top]$. Since $\mathcal{L}$ is clearly convex, we may find the optimal value by setting the gradient to 0:

$$\nabla\mathcal{L}(P) = 0 \iff P\mathbb{E}[ZZ^\top] = \mathbb{E}[XZ^\top].$$

We now make the hypothesis that $\mathbb{E}[ZZ^\top]$ is invertible. This means that the coordinates of Z are almost surely linearly independent, which is easy to guarantee in practice. With that assumption, we have

$$P = \mathbb{E}[XZ^\top]\mathbb{E}[ZZ^\top]^{-1}. \tag{2.4}$$

Therefore, if we denote the fitted values PZ by $\hat{X}$, we may write

$$X = \hat{X} + \eta, \tag{2.5}$$

where $\mathbb{E}[\hat{X}\eta^\top] = 0$, that is, the residual is orthogonal to the projection.

Now we go back to the structural equation, which, in the linear setting, is the following:

$$Y = X^\top\beta + \varepsilon. \tag{2.6}$$

If we substitute X using equation (2.5), we get

$$Y = \hat{X}^\top\beta + \eta^\top\beta + \varepsilon.$$

Multiply on the left by $\hat{X}$ and take expectations to obtain

$$\mathbb{E}[\hat{X}Y] = \mathbb{E}[\hat{X}\hat{X}^\top]\beta + \mathbb{E}[\hat{X}\eta^\top]\beta + \mathbb{E}[\hat{X}\varepsilon]. \tag{2.7}$$

We have already established that $\mathbb{E}[\hat{X}\eta^\top] = 0$. Notice also that

$$\mathbb{E}[\hat{X}\varepsilon] = \mathbb{E}[PZ\varepsilon] = \mathbb{E}[\mathbb{E}[PZ\varepsilon \mid Z]] = \mathbb{E}[PZ\mathbb{E}[\varepsilon \mid Z]] = 0,$$

by our definition of instrumental variable. Equation (2.7) then reduces to

$$\mathbb{E}[\hat{X}Y] = \mathbb{E}[\hat{X}\hat{X}^\top]\beta.$$

We would like to multiply both sides on the left by $\mathbb{E}[\hat{X}\hat{X}^\top]^{-1}$, but we must first check if this matrix is invertible. Expanding we have:

$$\mathbb{E}[\hat{X}\hat{X}^\top] = P\mathbb{E}[ZZ^\top]P^\top = \mathbb{E}[XZ^\top]\mathbb{E}[ZZ^\top]^{-1}\mathbb{E}[ZX^\top].$$

Therefore, if we require the rank of the matrix $\mathbb{E}[ZX^\top]$ to be d'_X, we have invertibility of $\mathbb{E}[\hat{X}\hat{X}^\top]$. Thus, we need to make two more assumptions: $d'_Z \geq d'_X$, which is equivalent to $d_Z \geq d_X$, and $\operatorname{rk}\mathbb{E}[ZX^\top] = d'_X$. The first assumption is a requirement of the second, and means that we need at least as many exogenous covariates as endogenous covariates in order to identify β. As for the second assumption, it is satisfied if X is sufficiently linearly related to Z (WOOLDRIDGE, 2001). Under these conditions, we have

$$\beta = \mathbb{E}[\hat{X}\hat{X}^\top]^{-1}\mathbb{E}[\hat{X}Y] \tag{2.8}$$

$$= \left[\mathbb{E}[XZ^\top]\mathbb{E}[ZZ^\top]^{-1}\mathbb{E}[ZX^\top]\right]^{-1}\mathbb{E}[XZ^\top]\mathbb{E}[ZZ^\top]^{-1}\mathbb{E}[ZY]. \tag{2.9}$$

The 2SLS estimator is then obtained by substituting the expectations by empirical versions, using the data in $\boldsymbol{X}, \boldsymbol{Z}$ and $\boldsymbol{Y}$. The analogue of expression (2.9) would be

$$\begin{aligned}\hat{\beta} = &\left[\left(\frac{1}{n}\sum_{i=1}^n X_i Z_i^\top\right)\left(\frac{1}{n}\sum_{i=1}^n Z_i Z_i^\top\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n Z_i X_i^\top\right)\right]^{-1} \\ &\cdot\left(\frac{1}{n}\sum_{i=1}^n X_i Z_i^\top\right)\left(\frac{1}{n}\sum_{i=1}^n Z_i Z_i^\top\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n Z_i Y_i\right).\end{aligned} \tag{2.10}$$

Notice that all n^{-1} factors cancel out, so this may be equivalently written as [1]

$$\begin{aligned}\hat{\beta} &= \left[\left(\sum_{i=1}^n X_i Z_i^\top\right)\left(\sum_{i=1}^n Z_i Z_i^\top\right)^{-1}\left(\sum_{i=1}^n Z_i X_i^\top\right)\right]^{-1} \\ &\quad\cdot\left(\sum_{i=1}^n X_i Z_i^\top\right)\left(\sum_{i=1}^n Z_i Z_i^\top\right)^{-1}\left(\sum_{i=1}^n Z_i Y_i\right) \\ &= \left[\boldsymbol{X}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{X}\right]^{-1}\boldsymbol{X}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{Y}.\end{aligned}$$

Letting $\hat{\boldsymbol{X}}$ denote $\boldsymbol{X}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$, we have

$$\hat{\beta} = \left(\hat{\boldsymbol{X}}\hat{\boldsymbol{X}}^\top\right)^{-1}\hat{\boldsymbol{X}}\boldsymbol{Y},$$

[1] We remind the reader that the transpose sign changes place when writing the 2SLS estimator using data matrices, since each observation of X and Z is a *line* in the corresponding matrix, not a column.

which is the empirical analogue of equation (2.8). This final form makes it clear that the estimator $\hat{\beta}$ is obtained by first performing one linear regression of $\boldsymbol{X}$ onto $\boldsymbol{Z}$, and then taking the fitted values $\hat{\boldsymbol{X}}$ and linearly regressing $\boldsymbol{Y}$ on them.

Using equation (2.10), together with the Law of Large Numbers and Slutsky's Theorem, one may prove the consistency of $\hat{\beta}$. Similar inspection allows one to establish asymptotic normality. For further theoretical properties of the 2SLS estimator, we refer the reader to (WOOLDRIDGE, 2001, Chapter 5), this section's main reference.

2.3.2 Numerical examples

We now provide numerical examples to strengthen our intuition about the differences between OLS and 2SLS. We present two scenarios with the same joint distribution for X and Z, in which both are one dimensional. The scenarios also share the same structural function, which we set to be $h^\star(x) = \beta_0 + \beta_1 x$, with $\beta_0 = 0.3$ and $\beta_1 = 0.7$. The difference between both experiments is the distribution of ε, which is mildly codependent with that of X in the first experiment, and strongly codependent in the second.

The data generating process we use for X and Z is the following:

$$\begin{aligned} \delta_i &\overset{\text{iid}}{\sim} \mathcal{N}(0,1), \quad i = 1,2; \\ Z &= \Phi(\delta_1); \\ X &= \Phi(\rho\delta_1 + \sqrt{1-\rho^2}\delta_2), \end{aligned}$$

where Φ is the cumulative distribution function of the standard normal and $\rho \in (0,1)$ is fixed at 0.8. For the first scenario, we generate ε as follows:

$$\begin{aligned} \delta_3 \sim \mathcal{N}(0,1), \quad \delta_3 \perp\!\!\!\perp (\delta_1, \delta_2); \\ \varepsilon = \sigma \cdot (\eta\delta_2 + \sqrt{1-\eta^2}\delta_3), \end{aligned}$$

where $\sigma > 0$ and $\eta \in (0,1)$ are set to be 0.1 and 0.6, respectively. For the second scenario, we introduce more interdependence between X and ε:

$$\varepsilon = \sigma \cdot \left(\left[\eta\delta_2 + \sqrt{1-\eta^2}\delta_3 \right] + C_b(\delta_2 - b)^+ - C_a(\delta_2 - a)^- \right).$$

Here, σ and η have the same values as before. The additional terms are $C_b = 6, b = 0.7, C_a = 2$ and $a = 0.3$. Finally, in both formulations we have $Y = h^\star(x) + \varepsilon$.

The results of the experiments are in Figure 2. We can see that in both of them the 2SLS estimate is closer to the true values of β_0 and β_1 than the OLS estimate. As expected, the OLS estimate does not take endogeneity into account and, therefore, incorporates some bias into the final estimates. This effect worsens as the endogeneity becomes stronger. An important observation, which is not visible in the figure, is that, as the number of observations grows, the OLS estimate drifts further from the true values, while the 2SLS estimate becomes closer to them.

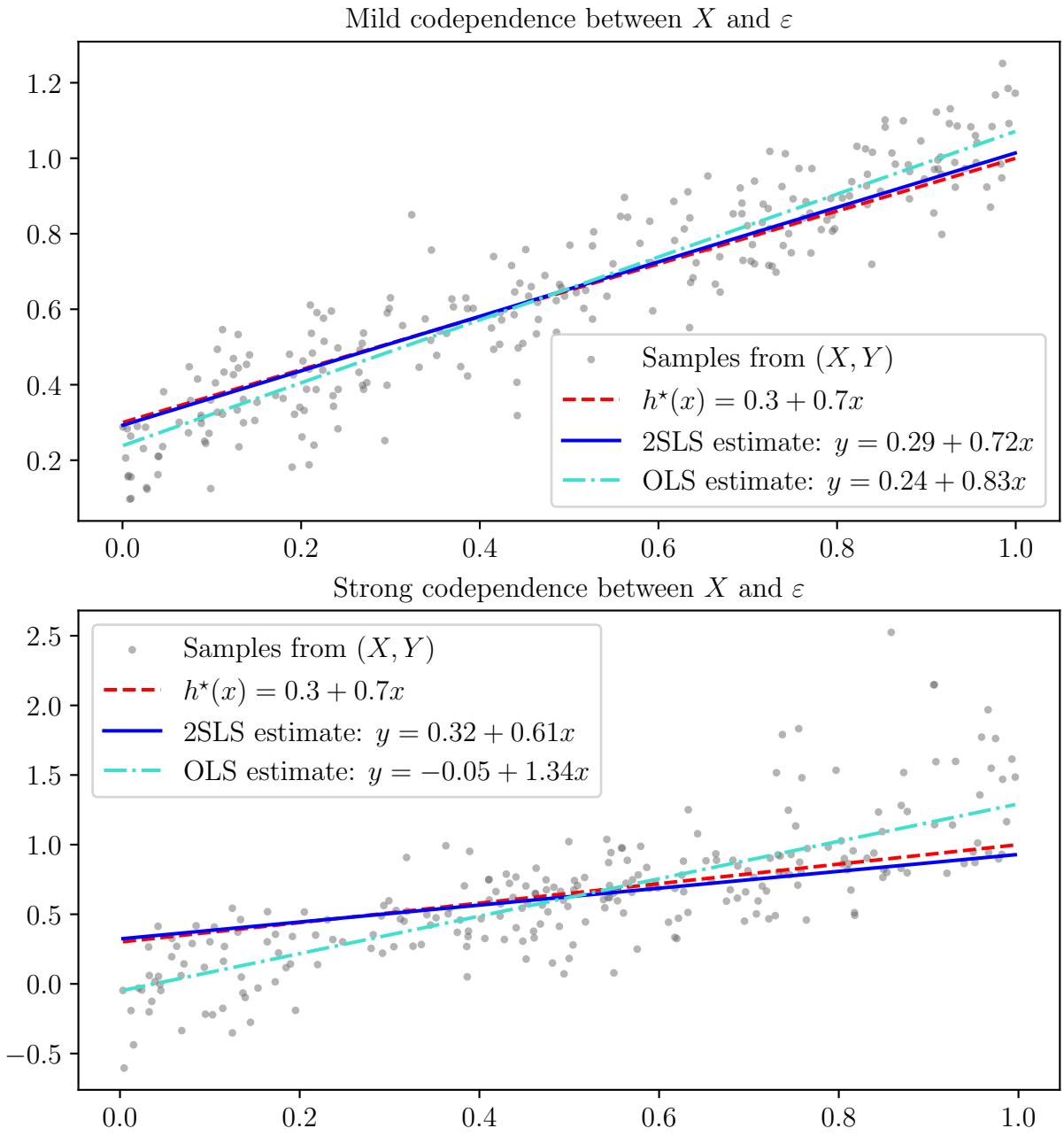

Figure 2 – Comparison between OLS and 2SLS under two different levels of endogeneity.

2.4 Nonparametric Instrumental Variable Regression

In nonparametric regression, we do not specify *a priori* a finite dimensional parametric form for the structural function (such as restricting it to be affine), and so we allow our search space to potentially be infinite dimensional. However, in doing this, we must still precisely define the infinite dimensional space where the solution will be searched for. Hence, we start by precisely defining the nonparametric regression problem given by (2.1)

2.4.1 Problem specification

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be the underlying probability space. Assume that $X : (\Omega, \mathcal{A}) \to (\mathbf{R}^{d_X}, \mathcal{B}(\mathbf{R}^{d_X}))$ and $\varepsilon : (\Omega, \mathcal{A}) \to (\mathbf{R}, \mathcal{B}(\mathbf{R}))$ are measurable[2] and, furthermore, that $\varepsilon \in L^1(\Omega, \mathcal{A}, \mathbb{P})$ with $\mathbb{E}[\varepsilon] = 0$. We also assume that $\mathbb{E}[\varepsilon \mid X]$ is *not* almost surely null and, hence, X is endogenous. Denote by $\mathbb{P}_X$ the distribution of the random variable[3] X, that is, the pushforward measure $\mathbb{P} \circ X^{-1}$ defined on $\mathcal{B}(\mathbf{R}^{d_X})$. We write $L^2(X)$ as a shorthand for the space $L^2(\mathbf{R}^{d_X}, \mathcal{B}(\mathbf{R}^{d_X}), \mathbb{P}_X)$ of real and square integrable (equivalence classes of) measurable functions defined on the measure space $(\mathbf{R}^{d_X}, \mathcal{B}(\mathbf{R}^{d_X}), \mathbb{P}_X)$. It's important to recall that the inner product and norm in $L^2(X)$ are given by $\langle h, g \rangle_{L^2(X)} = \mathbb{E}[h(X)g(X)]$ and $\|h\|^2_{L^2(X)} = \langle h, h \rangle_{L^2(X)} = \mathbb{E}[h(X)^2]$.

We assume there exists $h^\star \in L^2(X)$ such that (2.1) holds, that is, $Y = h^\star(X) + \varepsilon$. Finally, we assume there exists a random variable $Z : (\Omega, \mathcal{A}) \to (\mathbf{R}^{d_Z}, \mathcal{B}(\mathbf{R}^{d_Z}))$ such that Z qualifies as an instrumental variable, i.e., Z satisfies conditions 2.1 (i) and (ii). We define $\mathbb{P}_Z$ and $L^2(Z)$ in an manner analogous to $\mathbb{P}_X$ and $L^2(X)$. Our goal is to estimate $h^\star$ based on i.i.d. samples from the joint distribution of X, Z and Y.

2.4.2 Identification

An important question to ask after specifying the problem is whether the function $h^\star$ is identified. The answer is negative without further assumptions, which will be presented in this subsection. This discussion was inspired on (NEWEY; POWELL, 2003, Section 2).

Suppose there exists $\delta \in L^2(X)$ such that $\delta \neq 0$, but $\mathbb{E}[\delta(X) \mid Z] = 0$. Without loss of generality, we can assume[4] $\delta(X) \neq \mathbb{E}[\varepsilon \mid X]$. Defining $g \triangleq h^\star + \delta$ and $\eta \triangleq \varepsilon - \delta(X)$, we have

$$Y = g(X) + \eta,$$

where $\mathbb{E}[\eta \mid Z] = 0$ and $\mathbb{E}[\eta \mid X] \neq 0$. Hence, $g \neq h^\star$ but they are indistinguishable from the data generating process' perspective. Reciprocally, suppose that the only member of $L^2(X)$ which has null mean conditioned on Z is the null function. Then, given $g \in L^2(X)$ such that

$$Y = g(X) + \eta$$

with $\mathbb{E}[\eta \mid Z] = 0$, we have

$$0 = (g - h^\star)(X) + \eta - \varepsilon.$$

[2] We denote by $\mathcal{B}(\mathbf{R}^k)$ the Borel σ-algebra in $\mathbf{R}^k$.

[3] We use the term "random variable" when referring to scalar or vector valued measurable functions defined on $(\Omega, \mathcal{A})$.

[4] If it happens to be the case that $\mathbb{E}\left[\mathbb{E}[\varepsilon \mid X] \mid Z\right] = 0$ (which is *not* implied by our assumptions so far), we can simply take $\delta(X) = \lambda \mathbb{E}[\varepsilon \mid X]$, for some $\lambda \in \mathbf{R} \setminus \{0, 1\}$. Since, by hypothesis, $\mathbb{E}[\varepsilon \mid X] \neq 0$, this satisfies our requirements and is different from $\mathbb{E}[\varepsilon \mid X]$.

Conditioning on Z, we get

$$\mathbb{E}[(g - h^\star)(X) \mid Z] = 0,$$

which, by assumption, implies $h^\star = g$.

Therefore, a necessary and sufficient condition for identification of our regression problem is the following:

Assumption (Identification) If $\delta \in L^2(X)$ satisfies $\mathbb{E}[\delta(X) \mid Z] = 0$, then $\delta = 0$.

This condition has an interpretation in terms of the conditional expectation operator, which will be a key object in the construction of our estimator for $h^\star$. Let $h \in L^2(X)$. Notice that, by Jensen's inequality,

$$\mathbb{E}[(\mathbb{E}[h(X) \mid Z])^2] \leq \mathbb{E}[\mathbb{E}[h(X)^2 \mid Z]] = \mathbb{E}[h(X)^2] < +\infty. \tag{2.11}$$

Furthermore, since $\mathbb{E}[h(X) \mid Z]$ is a $\sigma(Z)$-measurable[5] random variable, by the Doob-Dynkin Lemma there exists a measurable function $f_h : (\mathbf{R}^{d_Z}, \mathcal{B}(\mathbf{R}^{d_Z})) \to (\mathbf{R}, \mathcal{B}(\mathbf{R}))$ such that

$$\mathbb{E}[h(X) \mid Z] = f_h(Z).$$

Under these conditions, we write $\mathbb{E}[h(X) \mid Z = z]$ for $f_h(z)$. The computation in (2.11) shows that $f_h \in L^2(Z)$ for every $h \in L^2(X)$. Therefore, we can define the operator $\mathcal{P} : L^2(X) \to L^2(Z)$ given by $\mathcal{P}[h] = f_h = \mathbb{E}[h(X) \mid Z = \cdot]$. This operator, called the *conditional expectation operator*, is clearly linear and, again by (2.11), also bounded, satisfying $\|\mathcal{P}\|_{\text{op}} \leq 1$. The identification assumption thus amounts to saying that the kernel of $\mathcal{P}$ is trivial, i.e., $\mathcal{P}$ is injective.

It is hard to quantify how restrictive this condition is for arbitrary X and Z, so we analyze it in a more familiar setting, the exponential family. We will use a classic completeness result for statistics in this family of distributions to reformulate the identification assumption in terms of more familiar objects. We first define completeness:

2.2 Definition (LEHMANN, 1959) We say that a family $\mathscr{P}$ of probability distributions on a measurable space $(E, \mathcal{E})$ is *complete* if

$$\int_E f(x)\ P(\mathrm{d}x) = 0 \quad \text{for all} \quad P \in \mathscr{P}$$

implies $f(x) = 0$ $\mathscr{P}$-a.e[6].

Then, a remarkable fact about the exponential family is the completeness of the natural statistics under a mild condition on the set of parameters:

[5] We denote by $\sigma(Z)$ the smallest σ-algebra in Ω with respect to which Z is measurable.

[6] We say that a statement $Q(x)$ is true $\mathscr{P}$-a.e. if there exists a set $N \in \mathcal{E}$ such that $Q(x)$ is true for $x \in E \setminus N$ and $P(N) = 0$ for all $P \in \mathscr{P}$.

2.3 Theorem (LEHMANN, 1959) Let Ξ be a subset of an Euclidian space with nonempty interior. Let X be a random vector with distribution P^θ parametrized by $\theta \in \Xi$ in the following manner:

$$P^\theta(\mathrm{d}x) = C(\theta) \exp\left\{\sum_{i=1}^{s} \theta_i T_i(x)\right\} \mu(\mathrm{d}x),$$

where μ is the underlying measure. Then, the family $\mathscr{P}_T$, formed by the distributions of the random vector $T(X) = (T_1(X), \ldots, T_s(X))$ as θ ranges through Ξ, is complete.

Using this result and under suitable hypotheses, we can reformulate the identification condition.

2.4 Theorem For $z \in \mathbf{R}^{d_z}$, let $\mathbb{Q}_z : \mathcal{B}(\mathbf{R}^{d_X}) \to [0,1]$ denote the conditional distribution of X given $Z = z$. Assume there exists $U \in \mathcal{B}(R^{d_Z})$ such that $\mathbb{P}_Z(U) = 1$ and for all $z \in U$ we have

$$\mathbb{Q}_z(\mathrm{d}x) = C(z) \exp(\alpha(z)^\top T(x))\ \mu(\mathrm{d}x)$$

for an underlying measure μ on $\mathcal{B}(\mathbf{R}^{d_X})$ and some functions $\alpha : \mathbf{R}^{d_z} \to \mathbf{R}^s$ and $T : \mathbf{R}^{d_x} \to \mathbf{R}^s$. Assume that T is injective and that the image of α restricted to U contains an open set. Then, $h^\star$ is identified.

Proof. Taking $\Xi = \alpha(U)$ and $\theta = \alpha(z)$, we see that the hypotheses of Theorem 2.3 are satisfied and, hence,

$$\mathscr{P} \triangleq \left\{\mathbb{Q}_z \circ T^{-1} : z \in U\right\}$$

is a complete family of probability distributions. Let $h \in L^2(X)$ be such that $\mathcal{P}[h] = 0$, i.e., $\mathbb{E}[h(X) \mid Z] = 0$ $\mathbb{P}_Z$-a.s. This means that the function

$$z \longmapsto \int_{\mathbf{R}^{d_x}} h(x)\ \mathbb{Q}_z(\mathrm{d}x)$$

is null $\mathbb{P}_Z$-a.s. Without loss of generality, we may assume that its null on U. But notice that, since T is injective, we can rewrite this integral as

$$\begin{aligned} 0 = \int_{\mathbf{R}^{d_x}} h(x)\ \mathbb{Q}_z(\mathrm{d}x) &= \int_{\mathbf{R}^{d_x}} (h \circ T^{-1})(T(x))\ \mathbb{Q}_z(\mathrm{d}x) \\ &= \int_{\mathbf{R}^s} (h \circ T^{-1})(t)\ (\mathbb{Q}_z \circ T^{-1})(\mathrm{d}t) \end{aligned}$$

for all $z \in U$ and some left inverse T^{-1} of T. By completeness of $\mathscr{P}$, this implies $h \circ\ T^{-1}(t) = 0$ $\mathscr{P}$-a.s. which, in turn, means that for all $z \in U$ we have

$$\begin{aligned} 1 = (\mathbb{Q}_z \circ T^{-1})[(h \circ T^{-1})(t) = 0] &= \mathbb{Q}_z[(h \circ T^{-1})(T(x)) = 0] \\ &= \mathbb{Q}_z[h(x) = 0]. \end{aligned}$$

Now, by the definition of conditional probability we have

$$\begin{aligned}\mathbb{P}_X[h(x) = 0] &= \int_{\mathbf{R}^{d_z}} \mathbb{Q}_z[h(x) = 0]\ \mathbb{P}_Z(\mathrm{d}z) \\ &= \int_U \mathbb{Q}_z[h(x) = 0]\ \mathbb{P}_Z(\mathrm{d}z) \\ &= 0.\end{aligned}$$

Therefore, h is the null function, which means $h^\star$ is identified. □

3 Conclusion

References

HERNÁN, Miguel A.; ROBINS, James M. **Causal Inference: What If**. [S.l.]: Chapman & Hall/CRC, 2020.

LEHMANN, E. L. **Testing Statistical Hypotheses**. [S.l.]: John Wiley & Sons, 1959.

NEWEY, Whitney K.; POWELL, James L. Instrumental Variable Estimation of Nonparametric Models. **Econometrica**, v. 71, n. 5, p. 1565–1578, 2003. DOI: http://dx.doi.org/10.1111/1468-0262.00459.

WOOLDRIDGE, Jeffrey M. **Econometric Analysis of Cross Section and Panel Data**. [S.l.]: The MIT Press, 2001. ISBN 9780262232197.

Appendix