FUNDAÇÃO GETULIO VARGAS

SCHOOL OF APPLIED MATHEMATICS

CAIO F. LINS PEIXOTO

NONPARAMETRIC INSTRUMENTAL VARIABLE REGRESSION THROUGH KERNEL METHODS AND STOCHASTIC GRADIENTS

Rio de Janeiro

2023

CAIO F. LINS PEIXOTO

NONPARAMETRIC INSTRUMENTAL VARIABLE REGRESSION THROUGH KERNEL METHODS AND STOCHASTIC GRADIENTS

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

CAIO F. LINS PEIXOTO

NONPARAMETRIC INSTRUMENTAL VARIABLE REGRESSION THROUGH KERNEL METHODS AND STOCHASTIC GRADIENTS

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

<div align="center">

—————————————————————

Yuri F. Saporito
School of Applied Mathematics

—————————————————————

Board Member 1
Institution 1

—————————————————————

Board Member 2
Institution 2

</div>

I dedicate this thesis to ...

Acknowledgements

Thanks, ...

" Bipedi bopedi bum! "
Albert Einstein

Abstract

Keywords:

Resumo

Palavras-chave:

List of Figures

List of Tables

Contents

1 Introduction

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

2 Instrumental Variable Regression

Our goal with this chapter is to provide an introduction to instrumental variables for mathematicians and statisticians unfamiliar with the topic. We assume throughout that all random variables are define on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$. By "random variable" we mean scalar or vector valued measurable functions with Ω as their domain.

> Rewrite chapter introduction.

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:

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

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 $\mathbb{E}[\eta \mid X, W] = 0$ a.s. and X and W are correlated. Hence, when we don't observe W and leave it to the error term, we end up estimating $h^\star(X) + \mathbb{E}[g^\star(W) \mid X]$. For example, if we want to regress a persons wage solely on her number of schooling years, there are other variables unaccounted for which influence both wages and schooling, such as natural ability. 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.

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.

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., $\mathbb{E}[\varepsilon \mid Z] = 0$.

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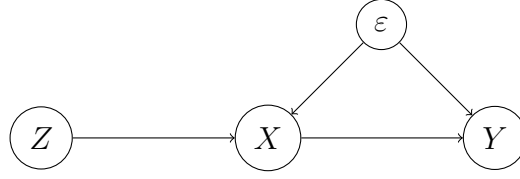
Figure 1 – Causal diagram for equation (2.1), where X is endogenous and Z is an IV.
Source: prepared by the author.

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:

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

and assume to have access to a random variable Z, taking values in $\mathbf{R}^{d_Z}$, satisfying conditions 2.1 (i) and (ii), so that Z is a valid instrumental variable. Hence, our data is composed of n independent joint samples $\{(X_i, Z_i, Y_i)\}_{i=1}^n$. Let $\boldsymbol{X} \in \mathbf{R}^{n \times (d_X + 1)}$ and $\boldsymbol{Z} \in \mathbf{R}^{n \times (d_Z + 1)}$ be the experiment design matrices with 1's in the first column.

Finish this section. Explain 2SLS and give examples of good IVs, as well as numerical examples.

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

We assume there exists an underlying probability triplet $(\Omega, \mathcal{A}, \mathbb{P})$. 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[1] and, fur-

thermore, 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 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 the introduction of (NEWEY; POWELL, 2003).

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[2] $\delta(X) \neq \mathbb{E}[\varepsilon \mid X]$. 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

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

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:

---

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

[2] 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 $\mathbb{E}[\varepsilon \mid X]$.

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

3 Conclusion

References

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

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